



Understanding and Managing Human-AI Interaction: Ethical Considerations

Teresa Hammerschmidt (née Heyder)

University of Bamberg

Bamberg 2026

Diese Arbeit hat der Fakultät Wirtschaftsinformatik und Angewandte Informatik als Dissertation vorgelegen.

Erstgutachter: Prof. Dr. Oliver Posegga

Zweitgutachter: Prof. Dr. Milad Mirbabaie

Mitglied der Promotionskommission: Prof. Dr. Mathias Klier

Datum der Abgabe: 23.10.2024

Datum der Disputation: 06.12.2024 (Tag der mündlichen Prüfung)

Dieses Werk ist als freie Onlineversion über das Forschungsinformationssystem (FIS; <https://fis.uni-bamberg.de>) der Universität Bamberg erreichbar.

Dieses Werk ist durch das deutsche Urheberrecht geschützt. Es steht Ihnen frei, dieses Werk auf jede Art und Weise zu nutzen, die durch die geltende Gesetzgebung zum deutschen Urheberrecht erlaubt ist. Für andere Verwendungszwecke müssen Sie die Erlaubnis der Rechteinhaberinnen und Rechteinhaber einholen.

URN: urn:nbn:de:bvb:473-irb-109667x

DOI: <https://doi.org/10.20378/irb-109667>

Von der genannten Lizenzangabe ausgenommen sind folgende Bestandteile dieser Dissertation:

Paper I "A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation" (S. X-CXXI), Paper VII "Navigating the Nexus of Ethical Standards and Moral Values" (S. CCXC-CCCXXII) und Paper IX "How Organizations Ethically Respond to Decision Dilemmas in AI Deployment" (S. CCCLXXVIII-CDXLIV) stehen unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Namensnennung 4.0
<https://creativecommons.org/licenses/by/4.0/>



Paper III "AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants?" (S. CXL-CXLIX), Paper VI "Ethical Management of Human-AI Interaction: A Theory Development Review" (S. CLXXVII-CCLXXXVIII), Paper VIII "Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI" (S. CCCXXIV-CCCLXXVI) und Paper X "How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making" (S. CDXLVI-CDLV) stehen unter der CC-Lizenz CC BY-NC-ND.

Lizenzvertrag: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0
<https://creativecommons.org/licenses/by-nc-nd/4.0/>



Dedicated to my sister

Elisabeth

Table of Contents

ACKNOWLEDGMENTS	III
ZUSAMMENFASSUNG (GERMAN SUMMARY)	V
1 INTRODUCTION	1
2 THEORETICAL FOUNDATIONS AND RELATED RESEARCH	9
2.1 The Role of Ethical Theories in Shaping Our Understanding of Ethical Human-AI Interaction	9
2.2 The Sociomaterial Notions Explaining Ethics and Moral Agency in Human-AI Interactions	18
3 METHODOLOGY	31
3.1 Systematic Literature Reviews (Primarily Deductive)	32
3.2 Empirical Investigations (Primarily Inductive)	34
4 MAIN RESEARCH RESULTS AND PAPER CONTRIBUTIONS	36
4.1 Understanding How Perceptions and Capabilities Influence Ethical Human-AI Interaction	39
Theoretical Foundations for RQ1: Paper I (Performativity)	39
<i>Paper I: Hammerschmidt, T.; Hafner, A.; Stolz, K.; Passlack, N.; Posegga, O. & Gerholz, K-H. (2025). A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation. Information Systems Frontiers (ISF), https://doi.org/10.1007/s10796-025-10596-0.</i>	39
Theoretical Foundations for RQ1: Paper II (Practices)	42
<i>Paper II: Hafner, A.; Hammerschmidt, T. & Stolz, K. (2024). Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review. Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany, https://aisel.aisnet.org/wi2024/46.</i>	42
Empirical Evidence for RQ1: Paper III (Relationality)	44
<i>Paper III: Hammerschmidt, T.; Passlack, N.; Stolz, K. & Posegga, O. (2024). AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants? Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, https://hdl.handle.net/10125/106426.</i>	44
Empirical Evidence for RQ1: Paper IV (Materiality)	46
<i>Paper IV: Hammerschmidt, T.; Stolz, K. & Posegga, O. (2024). How Leaders' Ambidexterity and Literacy Impact AI Capabilities of Organizations. Proceedings of 32nd the European Conference on Information Systems (ECIS), Paphos, Cyprus, https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14.</i>	46
Empirical Evidence for RQ1: Paper V (Inseparability)	48
<i>Paper V: Heyder, T. & Posegga (2021). Extending the Foundations of AI Literacy (Short Paper, Best Paper Nominee). Proceedings of the 46nd International Conference on Information Systems (ICIS), Austin, Texas, https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/.</i>	48

4.2 Understanding Ethical Management of Human-AI Interactions	52
Theoretical Foundations for RQ2: Paper VI (All Sociomaterial Notions in the Sociomaterial Cycle). 52	
<i>Paper VI: Heyder, T; Passlack, N. & Posegga, O. (2023) Ethical Management of Human-AI Interaction: A Theory Development Review. The Journal of Strategic Information Systems (JSIS), 32(3), 101772, https://doi.org/10.1016/j.jsis.2023.101772.</i>	52
Theoretical Foundations for RQ2: Paper VII (Practices).....	56
<i>Paper VII: Hammerschmidt, T. (2025). Navigating the Nexus of Ethical Standards and Moral Values. Ethics and Information Technology (EIT), 27(17), https://doi.org/10.1007/s10676-025-09826-5.</i>	56
Theoretical Foundations for RQ2: Paper VIII (Materiality).....	58
<i>Paper VIII: Hammerschmidt, T.; Stolz, K. & Posegga, O. (2025). Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI. Behaviour of Information Technology (BIT), https://doi.org/10.1080/0144929X.2025.2500451.</i>	58
Empirical Evidence for RQ2: Paper IX (Practices)	61
<i>Paper IX: Hammerschmidt, T.; Passlack, N. & Posegga, O. (working paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment. [previously 2nd revision round at JSIS].</i>	61
Empirical Evidence for RQ2: Paper X (Relationality)	63
<i>Paper X: Stolz, K.; Hammerschmidt T. & Posegga, O. (2024). How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making (Best Paper Nominee). Proceedings of 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, https://hdl.handle.net/10125/106970.</i>	63
 5 DISCUSSION	67
5.1 Theoretical Implications	67
5.2 Practical Implications	74
5.3 Methodological Implications	76
5.4 Limitations and Directions for Future Research.....	77
 6 CONCLUSION	78
 REFERENCES.....	81
 APPENDICES	VIII
Appendix I: Paper.....	VIII
Appendix II: Additional Publications.....	CDLVI

Acknowledgments

“Gratitude is not only the greatest of virtues, but the parent of all the others.” – Cicero

This dissertation would not have been possible without the immense support of my supervisor, PhD committee, colleagues, friends, and family. My gratitude runs deeper than words can express, and this journey has been as much theirs as it is mine. First, I want to extend my heartfelt thanks to my supervisors. Prof. Dr. Oliver Posegga, your unwavering guidance and willingness to let me explore my own pace has been invaluable. Thank you for giving me the freedom to test my ideas and for allowing my fast-track version of completing my papers and this thesis. I am also deeply grateful to Prof. Dr. Milad Mirbabaie and Prof. Dr. Mathias Klier for serving on my PhD committee. Your insightful feedback and enthusiasm for my work have been crucial, and I greatly appreciate your quick and uncomplicated support in navigating this journey with me. I also want to express my sincere appreciation to Prof. Dr. Kai Fischbach for offering me the opportunity to embark on my PhD at the University of Bamberg in the middle of the start of the corona pandemic lockdown and for your support throughout this journey. Special thanks to my old colleagues and supervisors at HUK-COBURG: Dirk Imer, Christiane Gerstlauer, Tiberius Andrei, Stefan Schuhman, and Stefanie Schmidt, for encouraging me to take this important step in my career and still feverish with me how it goes on.

To my incredible PhD co-authors, you have been pillars of support and inspiration. Nina Passlack, thank you for being a sounding board for my ideas, for your infectious enthusiasm during our last-minute brainstorming sessions, and for your endless encouragement—you are truly my positive potato. Dr. Katharina Stolz, it has been an honor to share this journey with you, from our master’s days to incredible experiences in places like the Yellow Mountains and Cyprus. Alina Hafner, Meike Bauer, and Prof. Dr. Karl-Heinz Gerholz—thank you for your ideas, collaboration, and shared wisdom along the way. I also want to extend my gratitude to the Department of Information Systems and Social Networks and its affiliate members. Special mention goes to Theresa Henn for your uplifting encouragement and friendship, Gudrun Stalkerich for being the heart and soul of the department, always knowing just what to do, Dr. Kathrin Eismann for helping with all thesis-related queries, and Dr. Diana Fischer-Preßler for your invaluable advice during our phone calls on preparing lectures and drafting papers. I would also like to thank Falco Klemm, Dr. Joscha Eirich, Scott Cooper, Dr. Peter Gloor and my incredible student assistants—Malou Schmidt, Exhaussé Matoko, Corinna Göllner, Nils Rahlmeier, Julia Halbritter, Chiara Seidenath, and Franziska Budde. Working with each of you has been a pleasure. Special thanks also go to the team of the project “Social Media Management for SMEs”, funded by the European Social Fund (ESF), which successfully transferred current research findings on social media and mobile business into practice for small and medium-sized enterprises in Bavaria (project duration 2020–2022). I am particularly grateful to Daniel Konadl, Katharina Schäfer-Siebert, and Torben Widmann, as well as Professors Julia Klier, Matthias Klier, Susanne Leist, Kai Fischbach, and Oliver Posegga, for their valuable contributions to the project’s success.

A huge thank you to the “ERBA Squad” of PhD students at the Faculty WIAI, University of Bamberg: Ferdinand Mittermeier, Julian Frey, Yannick Hildebrandt, Elias Grewe, Jan-Niklas Meckenstock, Leonie Fidler, Felix Haag, Marco Meier, Maximilian Raab, Carlo Stingl, Caroline Oelhorn, Sebastian Böhm, and many others—you know who you are. Your camaraderie made the highs more rewarding and the lows more bearable. I will miss the coffee breaks and hot-run evenings before conference submission deadlines as well as the shared excitement of whether we have made it one round further. To my new colleagues at Wiegand-Glas Holding GmbH, especially Lukas Neubauer—thank you for your understanding and flexibility during the whirlwind of finishing this thesis while starting a new, very existing, and lovely chapter in this wonderful company.

Finally, and most importantly, to my friends and family—none of this would have been possible without your unconditional love and support. To my family—Alexandra and Roman Heyder, Elisabeth Heyder and Ronny Lippert, Christoph and Stefanie Heyder, Peter and Helmi Wagner, Susanne, and in loving memory of Helmut Fehn-Wagner, as well as to my cousins and their families—thank you for your unwavering faith in me, for your constant love, and for helping me become who I am today. I am forever grateful. A special thank you to my parents, Alexandra and Roman, for instilling in me a deep trust in God and the faith that everything will work out as it should. Your guidance and belief have carried me through even the toughest times. To my sister, Elisabeth, thank you for taking care of our little farm with such dedication and heart. Knowing that everything was in your capable hands allowed me to focus on this dissertation. Your work and care behind the scenes mean more to me than words can say. To Markus Hammerschmidt—thank you for standing by me through these tough and crazy years. A special thanks to Markus’ family—Anita Hammerschmidt, Florian, Anna, and Magdalena Hammerschmidt, and Margot Kraus—for your constant encouragement and support. To my northern friends from my Kiel days—Marlena, Sureka, Mo, and especially Elena Kynast—you always brought good vibes and little surprises into my life, with Elena and little Anton making even Zoom wine sessions something to look forward to. To my school, bachelor’s, and master’s friends—specifically, Jenny Hofmann, Franzi Donde, Bine Böhnlein, and Sabrina Dormann with her family, especially my lovely godchild Klara—your friendship means the world to me. To my fellow members of the volunteer fire department and my dear hometown friends—thank you for keeping me grounded and reminding me of what truly matters in life. Special thanks to the Siebers, Browns, Wuschels, Kornes, Apels, Heinleins, and to Karina, Sascha, Thomas, Adi, Mo, Fritz, Roy, Maxi, and many more—you know who you are.

Cicero’s wisdom—“*Gratitude is not only the greatest of virtues, but the parent of all the others*”—perfectly encapsulates my journey through this dissertation. Every step of this process, from its challenges to its triumphs, has been made possible through the unwavering support, love, and faith of those around me. Gratitude has been the foundation. This work is not just a reflection of my efforts, but of the collective strength, kindness, and belief that others have shared with me. I will forever carry with me the lessons learned along the way, knowing that I have been truly blessed to have such incredible people in my life.

Zusammenfassung (German Summary)

Die vorliegende Dissertation beschäftigt sich mit dem Verständnis von Ethik, ethischen Implikationen und dem ethischen Management der Interaktion zwischen Menschen und Technologien, die auf Künstlicher Intelligenz (KI) beruhen. Die fortschreitenden selbstlernenden und (teil-)autonomen Fähigkeiten von KI-basierten Technologien führen zu sich verändernden Beziehungen zwischen Menschen und Technologien in organisationalen Kontexten (Berente et al., 2019; Dwivedi et al., 2021). Während bisherige digitale Technologien eher als Erweiterung menschlicher Intentionen fungierten, hat sich KI von einfachen automatisierten Technologien hin zu Systemen entwickelt, die eigenständig lernen, sich dynamisch anpassen und Entscheidungen treffen können (Rai et al., 2019; Raisch & Krakowski, 2021). Anders als frühere Technologien, die lediglich als passive Werkzeuge menschlicher Entscheidungen dienten, hat sich KI von automatisierten Systemen hin zu Technologien entwickelt, die in der Lage sind, selbstständig zu lernen, sich dynamisch anzupassen und autonome Entscheidungen zu treffen (Baird & Maruping, 2021; Endacott & Leonardi, 2022; Maedche et al., 2019). Das führt dazu, dass traditionelle Konzepte von “Human Agency” und “Material Agency”, die sich auf die Fähigkeit von menschlichen oder technologischen Akteuren beziehen, Handlungen auszuführen und auf die Umwelt einzuwirken (oft sog. “Handlungsfähigkeit”, “Handlungsmacht” oder “Akteurschaft”), hinterfragt werden (Hammerschmidt, Hafner, et al., 2025; Endacott & Leonardi, 2022). Die zunehmende Handlungsfähigkeit von KI verschiebt das Gleichgewicht von Autonomie und Verantwortung zwischen Menschen und Maschine, was Rollenbilder und Machtstrukturen innerhalb von Organisationen neu definiert (Constantiou et al., 2023; O’Neill et al., 2022). Diese Entwicklungen führen zu einer zunehmenden Verwischung der Grenze zwischen menschlicher Handlungsmacht und technologischer Autonomie (Endacott & Leonardi, 2022; O’Neill et al., 2022). Insbesondere, wenn Mitarbeitende beginnen, KI-basierte Technologien als gleichberechtigte Teammitglieder wahrzunehmen, wird es notwendig, die potenziell positiven, aber auch negativen ethischen Implikationen dieser Interaktionen für Individuen und Organisationen zu berücksichtigen (Baptista et al., 2020; Brendel et al., 2021; Heyder et al., 2023).

Im Gegensatz zu früheren Technologien agieren einige KI-basierten Technologien nicht nur als moralische Subjekte (sog. “Moral Patients”), die durch ethische Designprinzipien oder regulatorische Rahmenbedingungen beeinflusst werden, sondern übernehmen in gewissen Situationen auch eine aktive Rolle als moralische Akteure (sog. “Moral Agent”); (Hammerschmidt, Hafner, et al., 2025). Sie treffen Entscheidungen mit moralischen Implikationen für Menschen, basierend auf erlernten menschlichen Werten durch menschliche Dateneingaben (Allen et al., 2000; Behdadi & Munthe, 2020; Floridi & Sanders, 2004). Diese doppelte Rolle der KI als moralischer Akteur und moralisches Subjekt erweitert die traditionellen Vorstellungen von moralischer Verantwortung und schafft neue Herausforderungen, wenn es darum geht, zu definieren, was eine ethische Mensch-KI-Interaktionen bedeutet (Heyder et al., 2023). Dabei prägen die Dynamiken zwischen menschlichen Fähigkeiten (z. B. Wissen, Kompetenzen) und technologischen Fähigkeiten (z. B. objektive Ausgaben) die menschliche Wahrnehmung von KI

und deren moralische Tragweite, was wiederum das ethische Verhalten dieser Interaktion bestimmt. Beispielweise schafft die zunehmende Unabhängigkeit von KI-basierten Technologien moralische Pufferzonen für Menschen, die sich mit den Folgen von KI-Entscheidungen unwohl fühlen (Holford, 2022). Die erste Forschungsfrage lautet daher: *Wie beeinflussen menschliche Wahrnehmungen von KI sowie die Fähigkeiten sowohl von Menschen als auch von KI-basierten Technologien die ethische Mensch-KI-Interaktion?*

Organisationen stehen vor komplexen strategischen Dilemmata beim Einsatz von KI, die eine Abwägung zwischen den Vorteilen der Technologie und den damit verbundenen ethischen Risiken erfordern (Kegel & Ghanem, 2024; Shahid et al., 2023). Dies wird besonders dann relevant, wenn unerwartete Konsequenzen auftreten und die Zuweisung von Verantwortung – sei es an Entwickelnde, Nutzende oder Organisationen – unklar wird (Bogosian, 2017; Martin, 2019). Dabei hängt die moralische Bewertung stark von den unterschiedlichen Verständnissen von Ethik ab, was anhand bekannter, normativer ethischer Theorien – Tugendethik, Deontologie und Utilitarismus – aufgezeigt werden kann (Behdadi & Munthe, 2020; Brendel et al., 2021). Während die Tugendethik die Bedeutung persönlicher Motivation und ethischen Verhaltens in der Interaktion mit KI betont, fordert die Deontologie die Einhaltung universeller moralischer Prinzipien und der Utilitarismus konzentriert sich darauf, die besten Ergebnisse für die größtmögliche Anzahl von Menschen zu erzielen (Behdadi & Munthe, 2020; Brendel et al., 2021). Wenn verschiedene Personen in Organisationen mit unterschiedlichen ethischen Blickwinkeln zusammenkommen, um KI verantwortungsvoll zu designen und im Unternehmen zu integrieren, kann es zu ethischen Dilemmata kommen (Hammerschmidt, 2025). Daher müssen Organisationen Strategien entwickeln, um ethische Prinzipien mit moralischen Werten in organisatorischen Entscheidungsprozessen einbetten zu können (Berente et al., 2019; Mirbabaie et al., 2022). Die zweite Forschungsfrage lautet daher: *Wie können wir ethisches Management von Mensch-KI-Interaktionen verstehen?*

Durch die Kombination von systematischen Literaturanalysen mit qualitativ-empirischen Untersuchungen geht diese Dissertation den beiden genannten Forschungsfragen nach und bildet dabei einen neuen Ansatz von Theoriebildung nach Burton-Jones et al. (2021). Zu diesem Zweck baut diese Dissertation auf der Konzeptualisierung von Heyder et al. (2023) auf, die Kernkonzepte von Soziomaterialität nach Orlikowski (2007, 2010) mit ethischen Theorien kombiniert, um unterschiedliche ontologische Positionen zu Ethik in der Interaktion zwischen Menschen und KI zu erklären. Dadurch kann diese Dissertation besser erklären, wie ethische Theorien das Verständnis von moralischer Handlungsmacht in der KI beeinflussen und wie organisatorische Praktiken ethische Bedenken sowie ethisches Verhalten in der Mensch-KI Interaktion entweder mindern oder verschärfen können.

Die Dissertation zeigt Grenzen auf, die sich aus der isolierten Anwendung ethischer Theorien bei dem ethischen Management der Mensch-KI Interaktion ergeben. Sie plädiert für einen differenzierteren Ansatz, der mögliche Konflikte zwischen den moralischen Maximen verschiedener Menschen und den ethischen Prinzipien von KI berücksichtigt. Die Ergebnisse zeigen, dass ethisches Management ein

Gleichgewicht zwischen menschlicher Aufsicht und den wachsenden Entscheidungsfähigkeiten von KI erfordert. Es muss sichergestellt werden, dass KI-basierte Technologien mit soliden ethischen Richtlinien entwickelt werden und dass die Nutzenden durch Wissen und Kompetenzen in der Lage sind, KI-Ergebnisse kritisch zu bewerten und gegebenenfalls einzugreifen. Zudem müssen sich Organisationen der veränderten Rollen in Hinblick auf die Unterscheidung menschlicher und KI-basierter Handlungsfähigkeit bewusst werden, um eine Fehlzuschreibung moralischer Verantwortung auf die Technologie sowie mögliche Rollenkonflikte zu vermeiden.

Die Dissertation leistet einen Beitrag zur wachsenden Forschung über Ethik in der Mensch-KI-Interaktion. Zunächst reagiert sie auf die Forderungen nach einer allgemeinen Theorie (z. B. Berente et al., 2019; Hagendorff, 2020; Mirbabaie et al., 2022; Siau and Wang, 2020), die die Komplexitäten von Ethik in der Mensch-KI Interaktion adressiert, indem normative ethische Theorien mit dem Konzept von Soziomaterialität kombiniert werden. Dies ermöglicht ein besseres Verständnis von KI und menschlicher Handlungsfähigkeit in der Mensch-KI-Interaktion. Zudem erweitert sie den Diskurs über Soziomaterialität, indem sie ein stärkeres Verständnis zur Handlungsfähigkeit von KI-basierten Technologien im Vergleich zum Menschen schafft, in dem beide ontologischen Positionen (relationale sowie substanzielle Positionen) bei der Konzeptualisierung berücksichtigt werden. Dabei fordert die Dissertation eine Neubewertung des Verständnisses der Konzepte von Handlungsmacht und Ethik, da die Grenzen zwischen menschlicher Intention und der Autonomie der Technologie verschwimmen. Ferner vertieft die Dissertation das Wissen darüber, wie unterschiedliche Interessensgruppen verschiedene Strategien zur Steuerung der Mensch-KI Interaktion verfolgen. Die Dissertation zeigt auf, wie die Förderung von Wissen und Kompetenz rund um KI moralische Ungleichheiten in diesen Interaktionen verringern können. Ethisches Management umfasst daher nicht nur das Design von KI auf der Grundlage ethischer Prinzipien, sondern auch die Förderung einer ethisch bewussten und geschulten Gesellschaft, die diese Technologien verantwortungsvoll nutzen kann. Diese Dissertation legt den Grundstein für zukünftige Studien, die die komplexen Beziehungen zwischen ontologischen Positionen, ethischen Theorien und Soziomaterialität vertiefen.

Die Dissertation ist wie folgt aufgebaut: Zunächst werden ethischer Theorien (Abschnitt 2.1) sowie die Kernkonzepte von Soziomaterialität (Abschnitt 2.2) vorgestellt. Abschnitt 3 beschreibt die Methodik, die deduktive Literaturrecherchen (Abschnitt 3.1) mit induktiven empirischen Untersuchungen (Abschnitt 3.2) kombiniert. Abschnitt 4 präsentiert die Ergebnisse der insgesamt zehn Artikel, aufgeteilt in zwei Teile: Wie Wahrnehmungen und Fähigkeiten ethische Mensch-KI-Interaktion beeinflussen (Abschnitt 4.1) und das ethische Management dieser Interaktionen (Abschnitt 4.2). Abschnitt 5 erörtert die theoretischen (5.1), praktischen (5.2) und methodischen Implikationen (5.3) und gibt Ausblicke auf zukünftige Forschung anhand von Limitationen (5.4). Die Arbeit endet mit einem Fazit.

1 Introduction

The increasing self-learning and (semi-)autonomous capabilities of artificial intelligence (AI)¹ technologies are transforming human-AI interactions in organizations, challenging traditional concepts of agency and raising critical ethical concerns about how these interactions should be managed (Berente et al., 2019; Dwivedi et al., 2021; Endacott & Leonardi, 2022). Unlike earlier technologies, which functioned merely as extensions of human intent, AI has evolved from technologies of limited, programmed automation to a set of technologies capable of learning, adapting, and making decisions autonomously (Rai et al., 2019; Raisch & Krakowski, 2021). AI systems no longer simply react to human commands with a static and passive response but also actively shape interactions with dynamic and more interactive responses that are not always predictable, beginning to assume responsibility for tasks independently (Baird & Maruping, 2021; Endacott & Leonardi, 2022; Maedche et al., 2019). As AI's agency develops, the balance of autonomy between humans and AI is shifting, transforming business roles and reshaping organizational power structures (Constantiou et al., 2023; O'Neill et al., 2022; Strich et al., 2021). This development increasingly blurs the boundaries between human agency and technological autonomy (Endacott & Leonardi, 2022; O'Neill et al., 2022). When employees begin to perceive AI as a genuine team member, organizations must recognize that these dynamics can simultaneously generate both positive and negative ethical implications for individuals and the organizations (Baptista et al., 2020; Brendel et al., 2021; Marabelli & Galliers, 2017). As a result, organizations face complex decision dilemmas in their AI deployment and utilization strategies, requiring a balance between harnessing AI's benefits and mitigating its ethical risks (Kegel & Ghanem, 2024; Shahid et al., 2023). Ethical management of human-AI interaction, therefore, demands a reassessment of organizational strategies and practices to address the ethical concerns accompanying AI's changing role in the workplace (Baptista et al., 2020; Berente et al., 2019; Brendel et al., 2021).

Additionally, given the advances in AI's agency, understanding the interplay between AI and human moral agency becomes critical for managing human-AI interaction ethically (Constantinescu et al., 2021; Cunneen et al., 2019; Heyder et al., 2023). Unlike previous technologies, some AI technologies function both as moral patients, being affected and constrained by the moral actions of others (e.g., ethical design principles, regulatory frameworks) (Allen et al., 2000). At the same time, they can also function as moral agents, possessing a degree of moral decision-making autonomy by executing tasks with moral implications for people based on self-learning from human moral values (Allen et al., 2000; Behdadi & Munthe, 2020; Floridi & Sanders, 2004). This dual role extends the traditional boundaries

¹ This dissertation adopts a broad definition of AI technologies, acknowledging that findings related to one specific type of AI may not be directly transferable to others. To address diverse normative ethical perspectives and recognize context-specific ontologies that create the agency of various AI technologies, this work integrates and examines insights from multiple AI technologies across different application domains and ethical contexts. A narrowly focused discussion might overlook essential aspects of the ethical management of AI design, deployment, use, and outcomes. Consequently, AI technologies in this study are broadly defined as technological artifacts incorporating systems with cognitive capabilities—such as reasoning, learning from experience based on human data input, problem-solving, and natural language understanding—that are typically associated with human intelligence (Rai et al., 2019). The scope includes several AI technologies such as autonomous vehicles, generative AI tools, algorithmically managed platforms, crime detection systems, medical AI assistants, recruiting agents, and managerial decision-making systems, with attention given to specific ethical considerations where relevant.

of moral involvement, as AI's capacity to learn from human data embeds a form of morality within the technology that surpasses the scope of ethical design principles alone (Etzioni & Etzioni, 2017; Floridi & Sanders, 2004). Consequently, this extension creates a new form of artificial moral agency that is distinct from traditional forms of human moral agency (Allen et al., 2000; Behdadi & Munthe, 2020) and that increasingly blurs the lines of accountability, particularly when unintended consequences arise (Giermindl et al., 2022; Horneber & Laumer, 2023). For instance, when errors or harms occur, determining whether developers, users, or organizations are accountable becomes challenging, leading to moral uncertainties and ambiguous responsibilities (Bogosian, 2017; Martin, 2019). Additionally, AI's growing independence creates "moral buffers" that distance humans from the outcomes of AI decisions (Holford, 2022, p. 148). This raises critical questions such as "where duty falls when algorithms have varied degrees of agency to act" (Pansoni et al., 2023, p. 1152), which necessitate a rethinking of what comprises the ethical management of human-AI interactions within organizations (Baptista et al., 2020; Berente et al., 2019; Brendel et al., 2021). Therefore, the ethical management of human-AI interaction in organizations requires not only establishing clear governance frameworks and principles, considering AI as a moral patient, but also recognizing the broader cultural and behavioral impacts that arise when AI functions as a moral agent (Cooper et al., 2022; Peters et al., 2020).

This shift requires reconsidering previous conceptualizations of the interdependencies between human and AI agency, challenging traditional assumptions about sociomateriality. Sociomateriality, a theoretical lens, views technology as a "material" entity intertwined with "social" phenomena such as culture, meanings, and human decisions (Orlikowski, 2007, 2010; Orlikowski & Scott, 2008). It explains how power dynamics in the interrelations between social and material entities can shape ethical and nonethical behaviors in everyday human-technology interactions (Cecez-Kecmanovic et al., 2014). Scott and Orlikowski (2013) critiqued early sociomaterial research for not adequately accounting for the role of human agency in shaping technology's design and use. Current critiques shift attention to the role of AI's agency in transforming the nature of human-AI interaction, challenging traditional views of technologies as neutral entities that only respond to human interactions (Endacott & Leonardi, 2022; Hammerschmidt, 2025; Heyder et al., 2023). Cecek-Kecmanovic (2014) described this shift as the non-neutral co-constitution of material agency. Non-neutrality here refers to the idea that some AI systems introduce moral considerations into the human-AI interaction that are not entirely controllable or predictable by humans, thereby influencing ethical behavior and outcomes as moral agents (Allen et al., 2000; Behdadi & Munthe, 2020; Floridi & Sanders, 2004). Co-constitution implies that ethical behavior and outcomes are mutually shaped by both the human and technological entities within the human-AI interaction, which distributes moral agency across both human and technological actors (e.g., Behdadi & Munthe, 2020; Endacott & Leonardi, 2022; Floridi & Sanders, 2004; Niederman & Baker, 2023; Townsend et al., 2022). More distinct investigations are required to account for this non-neutral co-constitution of moral agency between technologies and social actors, as this shapes the ethical behavior and outcomes of human-AI interactions that we do not fully understand yet (Heyder et al., 2023). Thus,

expanding the sociomaterial discourse to better capture the shifting dynamics of AI and human morality, and their implications for ethical human-AI interaction, is essential for deepening our understanding of how to manage these interactions ethically

Before addressing how to manage ethical human-AI interactions, we must first deepen our understanding of the factors that influence these interactions. To better capture how the shifting dynamics of moral agency between humans and AI affect ethical human-AI interaction, we need to be aware of how human perceptions of AI, along with the capabilities of both humans and AI, shape each other (or do not) and how this affects the ethical nature of their interactions. Thereby, it is important to consider the different ontological positions that shape our interpretation of the entanglement between human and AI moral agency, as this depends on AI capabilities and the perceptions that this creates among people (Heyder et al., 2023). A substantial ontological position suggests that the capabilities of humans and AI, as well as human perceptions of AI, are independent and stable, while a relational ontological position views human and AI capabilities as mutually dependent (Hammerschmidt, 2025; Heyder et al., 2023). So far, much of the sociomaterial discourse has focused on comparing these ontological positions rather than integrating them to create distinct but potentially equally valid insights (Cecez-Kecmanovic et al., 2014; M. Jones, 2014). However, both ontological positions are essential when examining ethical human-AI interactions (Heyder et al., 2023). For instance, transparent AI systems, which allow users to understand how algorithms function, might foster ethical interactions by promoting oversight (Asatiani et al., 2020; Bleher & Braun, 2023). However, greater transparency can also lead to opportunistic behavior, where users manipulate AI systems based on perceived unfairness (M. K. Lee, 2018; Liu et al., 2021). This example illustrates relational ontology, where the ethicality of the interaction is shaped by the interplay between AI's capabilities to provide explanations and human users' perceptions of fairness. In contrast, black-box AI, which lacks transparency and explanations, might be seen as leading to unethical interactions due to limited understanding (Asatiani et al., 2020; Rai, 2020). However, this opacity may also prompt users to apply their own ethical reasoning when justifying AI's decisions (Endacott & Leonardi, 2022), reflecting substantial ontology, where human ethical capabilities exist independently of the technology's missing explanation capabilities. In both examples, the interaction between AI and human capabilities, as well as human perceptions of AI, plays a key role in determining whether human-AI interactions are ethical (e.g., Hammerschmidt et al., 2024; Heyder & Posegga, 2021). Whether viewed through a relational ontology, which highlights the dynamic interplay between human and AI capabilities, or a substantial ontology, which treats human ethics as independent, these perspectives are essential to understanding ethical human-AI interaction. Therefore, this dissertation will incorporate both ontological positions to explore the complex interplay between social and material capabilities and human perceptions in shaping human and AI moral agency, as this is a crucial step toward understanding ethical human-AI interactions before attempting to manage them ethically. Thus, the first research question (RQ1) is: **How do human perceptions and the capabilities of both humans and AI influence ethical human-AI interaction?**

Besides analyzing how perceptions and capabilities influence ethical human-AI interaction, it is further crucial to increase our understanding of how the ethical management of human-AI interaction can be based on different strategies, as scholars such as Berente et al. (2019), Mirbabaie et al. (2022), Hagedorff (2020), and Siau and Wang (2020) argue that there is a lack of general theory that considers various ethical perspectives when describing the complexity of ethics in human-AI interaction. In order to understand the necessity of addressing ethics within human-AI interactions, a clear distinction between ethics and morality is required. Ethics, broadly defined, is “the capacity to think critically about moral values and direct our actions in terms of such values” (Churchill, 1999, p. 259). At the core of morality, societal moral values and cultural norms of specific “moral communities” create different ontological realities of “good” or “bad” behavior (Champagne & Tonkens, 2023, p. 27:8; Søvik, 2022). Thus, the ethical evaluation of actions as “right” or “wrong” depends on normative ethical perspectives shaped by the morality of cultural contexts (Hammerschmidt, Hafner, et al., 2025). Thus, our understanding of ethics and morality, like our understanding of AI and human moral agency, is influenced by different ontological positions (Cunneen et al., 2019; Kieslich et al., 2022). For example, a relational ontological position would suggest that the ethical behavior of both humans and AI is shaped through interactions, whereas a substantial position would treat AI technologies and humans as distinct entities with inherent virtues (humans) and ethical principles (AI) that remain unchanged by their interactions (Heyder et al., 2023). These varying ontological positions can lead to divergent views on the ethical obligations of social and material entities, affecting the behavior of various stakeholders involved in human-AI interactions (Cunneen et al., 2019; Kieslich et al., 2022). Consequently, ontological frameworks influence the ethical assessment of right and wrong, shaping different managerial strategies regarding the ethical design, deployment, and use of AI technologies, which may potentially clash when grounded in different ethical perspectives (Hammerschmidt, 2025).

Deontology, virtue ethics, and utilitarianism are prominent normative ethical perspectives that offer various ontological frameworks for understanding ethics in human-AI interactions (Bankins & Formosa, 2023; Bilal et al., 2021; Vanhée & Borit, 2022). *Utilitarianism*, a consequentialist ethical theory, evaluates the morality of actions based on their outcomes, aiming to maximize overall happiness (Anderson & Anderson, 2007; Yu et al., 2021). In organizational contexts, AI strategies often adopt a utilitarian perspective, prioritizing the greatest benefits for the majority of stakeholders (Emdad et al., 2023; Morton et al., 2022). However, utilitarianism faces criticism for reducing morality to a cost-benefit analysis, potentially neglecting minority interests (Serafimova, 2020). In contrast, *deontology*, derived from Kant’s categorical imperative, emphasizes that individuals should adhere to ethical principles that they would want to become a universal law (Kant, 1991). This theory holds that certain actions are inherently moral or immoral, regardless of their outcomes (Gaus, 2001; Unver, 2023). Deontology is often referred to as duty ethics because it emphasizes the role of principles, rules, norms, and standards in guiding ethical behavior according to obligations (Bowie, 1998). This perspective is prevalent in regulated environments where ethical behavior is governed by law, such as the European AI Act, which

seeks to ensure transparency, privacy, and fairness for all users equally (H. Roberts et al., 2023; Stahl et al., 2022). However, deontological frameworks shaped by dominant nations may oversimplify or inadequately address complex global ethical issues (Millar, 2016; Swanepoel, 2021). *Virtue ethics*, on the other hand, centers on individual moral values and sensibilities, allowing for divergence from established standards based on personal moral assumptions (Bogosian, 2017; Coeckelbergh, 2020). It emphasizes the cultivation of virtues that form the foundation of human moral agency based on free will (Coeckelbergh, 2020; Puri, 2020; Shneiderman, 2020). Although AI can learn to mimic moral behavior from training data, its moral agency remains restricted (Cook, 2023; Etzioni & Etzioni, 2017; Moor, 2006; Saariluoma & Leikas, 2020). Each of these ethical perspectives offers distinct insights into how ethical behavior and outcomes can be assessed. However, relying solely on one framework does not capture the full complexity of ethical issues in human-AI interactions. As a result, the primary challenge for the ethical management of human-AI interaction lies in aligning these diverse perspectives, as their conflicting maxims can lead to moral dilemmas. For instance, deontological principles may establish an ethical baseline for AI design but could clash with the personal moral values central to virtue ethics (Grønsund & Aanestad, 2020; Madaio et al., 2020). Similarly, utilitarian strategies may benefit the majority of stakeholders but, in doing so, might disadvantage minority groups, raising concerns that conflict with Kantian principles of universal moral law (Buruk et al., 2020; Fukuda-Parr & Gibbons, 2021; Meijer et al., 2023). These conflicts highlight the moral dilemmas that arise when different ethical maxims, societal values, and principles intersect (McDonald & Pan, 2020; Shank & Gott, 2020). In practice, these dilemmas are especially intricate, as neither individuals nor organizations strictly adhere to a single ethical perspective (B. Zhang et al., 2021), leading to misaligned moral values and competing ethical frameworks, which complicate the formulation of ethical AI strategies (Brendel et al., 2021; Mirbabaie et al., 2022). Therefore, managing human-AI interactions ethically necessitates the alignment of various ethical theories, each emerging from different ontological environments and shaping distinct maxims for ethical behavior and moral outcomes. This dissertation aims to consider the interplay of these different ethical perspectives when answering the second research question (RQ2): **How can we understand the ethical management of human-AI interaction?**

The primary objective of this dissertation is to investigate the understanding and management of human-AI interaction from an ethical perspective, accounting for the changing dynamics between human and AI moral agency. This dissertation does not focus on managing ethics in a general sense, as ethics itself cannot be directly managed. Rather, it explores how ethical considerations in human-AI interaction can be integrated into managerial decision-making and strategies, helping to overcome potential moral decision dilemmas. The central aim is to examine how ethically aligned management practices can facilitate more ethical outcomes and behaviors in these interactions. Thus, the focus is on explaining what makes human-AI interaction ethical and how management can play a pivotal role in promoting these ethical dimensions. Thereby, this dissertation analyzes how different ontological positions shape our understanding of ethics and moral agency in human-AI interaction, and how

organizations can manage these interactions ethically. By examining the theoretical and practical challenges arising from AI's changing agency, this dissertation expands on previous conceptualizations of ethics, (moral) agency, and sociomateriality. Therefore, this dissertation investigates well-established ethical theories through the lens of sociomateriality to explain the ethical interdependencies between AI and human moral agency, particularly how these interactions influence ethical or unethical behavior and outcomes based on varying ontological perspectives. It integrates established normative ethical theories, such as deontology, utilitarianism, and virtue ethics, to demonstrate how different ontological contexts shape perceptions of ethics and moral agency. By doing so, this dissertation contributes to a broader understanding of ethics when different human stakeholders (such as developers, regulators, managers, and users) regulate, design, manage, implement, and utilize AI technologies throughout their life cycle. This further helps describe how human-AI interaction can be managed ethically, accounting for various ethical concerns and moral dilemmas that arise when formulating AI strategies within organizations.

To address these goals, the methodological approach of this dissertation uses theoretical foundations derived from systematic literature reviews and enhances them with empirical insights on perceptions of AI and human moral agency to explain “new phenomena emerging in a changing world” and to extend existing theories (Burton-Jones et al., 2021, p. 303). To do so, this dissertation builds on the conceptualization by Heyder et al. (2023), which integrates the weak and strong notions of sociomateriality following Jones (2014)—inseparability, relationality, performativity, materiality, and practices—into Leonardi's (2012) conceptualization of sociotechnical systems, which melds the discourse of sociomateriality within organizations as sociotechnical systems, to explain different ontological positions on ethics and moral agency within human-AI interaction. Both are grounded in the discourse on sociomateriality advanced by Orlikowski (2007, 2010) and Orlikowski and Scott (2008). By leveraging the weak and strong notions of sociomateriality, this dissertation can provide a more nuanced analysis of the different ethical interdependencies between AI and human agency, demonstrating how AI abilities reshape relationships, capabilities, and perceptions among various human stakeholders who engage with AI at different stages of the AI life cycle. This adopts the so-called extend strategy of “Next-Generation Theorizing” proposed by Burton-Jones et al. (2021, p. 303).

This dissertation aims to make five contributions to information systems (IS) research. *First*, it responds to the calls from Berente et al. (2019), Hagendorff (2020), Mirbabaie et al. (2022), and Siau and Wang (2020) for a general theory that addresses the complexities of ethics in AI transformation. By incorporating normative ethical theories into the lens of sociomateriality, this dissertation offers insights into how ethical considerations can guide the understanding of AI and human (moral) agency in managing human-AI interactions, which particularly addresses RQ1. *Second*, this dissertation expands the discourse on sociomateriality by deeply examining changes in the interplay between human and AI (moral) agency through the distinct weak and strong notions of sociomateriality, integrating different ontological positions, which particularly address RQ1. In doing so, this dissertation responds to the call of Cecez-Kecmanovic (2014, p. 826) to examine the “non-neutral (i.e., knowledge/power)

performativity of social and technological entanglements,” and by integrating the ethical perspective and practice-related discussions on moral agency, we add novel insights to existing sociomateriality conceptualizations, as requested by Jones (2014). *Third*, this dissertation addresses the calls of Alsheibani et al. (2020), Vial (2019), and Zhang et al. (2022) to deepen knowledge of why and how different stakeholders adopt diverse strategic approaches to manage human-AI interaction by dealing with potential decision dilemmas ethically. This includes considering their different ontological and normative ethical assumptions guiding their approaches to overcome those conflicts and particularly addresses RQ2. *Fourth*, this dissertation contributes to research on how fostering AI literacy can help mitigate moral inequalities in human-AI interactions, as emphasized by Carter et al. (2020), Choudrie et al. (2022), Schneider et al. (2023), and Vassilakopoulou and Hustad (2023). This involves exploring how managerial strategies can enhance an individual’s abilities to appropriately engage with AI, thereby bridging the divide between those who are able to benefit from AI technologies and those who are not, which particularly addresses RQ2. *Fifth*, this dissertation illustrates how “Next-Generation Theorizing” of Burton-Jones et al. (2021, p. 303) may be conducted by combining mainly deductive insights from systematic literature reviews building upon well-established research (Paré et al., 2015; Templier & Paré, 2018) with explorative and inductive insights from empirical investigations (Sarker et al., 2018) to combine previous ethical theories and sociomaterial conceptualizations to study new phenomena appearing in the IS field. Thereby, this dissertation’s nature of theory is explanation (Gregor, 2006), seeing theory as an intellectual conversation and a lens that is itself performative (Burton-Jones et al., 2021).

The structure of this dissertation is as follows: The next section presents the theoretical foundations and related research on the role of ethical theories in shaping our understanding of ethical human-AI interaction (Section 2.1) and the weak and strong notions of sociomateriality as a lens for analyzing ethics and moral agency based on different ontological positions (Section 2.2). The third section outlines the methodology, adopting the *extend strategy* of “Next-Generation Theorizing” by Burton-Jones et al. (2021, p. 303), combining primarily deductive systematic literature reviews (Section 3.1) with inductive empirical investigations (Section 3.2). The fourth section presents the results, divided into two parts: one focusing on how perceptions and capabilities influence ethical human-AI interaction, corresponding to the first research question (Section 4.1), including **Papers I to V**, and another focusing on the ethical management of human-AI interaction (Section 4.2), corresponding to the second research question and including **Papers VI to X** (see Table 1). For both subsections of Section 4, the dissertation’s papers are divided into those articles providing theoretical foundations based on systematic literature reviews and those providing empirical evidence based on field studies. An overview of the dissertation’s articles is given in Table 1.

Number & Topic	Contribution Type	Authors	Year & Status	Title	Publication Outlet & Ranking & DOI
Paper I (Morality)	Theoretical Foundations	Hammerschmidt , T.; Hafner, A.; Stolz, K.; Passlack, N., Posegga, O. & Gerholz, K-H.	2025 (published)	A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation	Information Systems Frontiers (ISF), B-Ranking (VHB JOURQUAL-4) https://doi.org/10.1007/s10796-025-10596-0
Paper II (Ethics)		Hafner, A.; Hammerschmidt , T. & Stolz, K.	2024 (published)	Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review.	Internationale Tagung Wirtschaftsinformatik (WI), B-Ranking (VHB JOURQUAL-4) https://aisel.aisnet.org/wi2024/46
Paper III (Perceptions)	Empirical Evidence	Hammerschmidt , T.; Passlack, N.; Stolz, K. & Posegga, O.	2024 (published)	AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants?	Hawaii International Conference on System Sciences (HICSS), B-Ranking (VHB JOURQUAL-4) https://aisel.aisnet.org/hicss-57/cl/machines_as_teammates/7/
Paper IV (Capabilities)		Hammerschmidt , T.; Stolz, K. & Posegga, O.	2024 (published)	How Leaders' Ambidexterity and Literacy Impact AI Capabilities of Organizations	European Conference on Information Systems (ECIS), A-Ranking (VHB JOURQUAL-4) https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14/
Paper V (Literacy)		Heyder , T. & Posegga, O.	2021 (published) Short Paper, Best Paper Nominee	Extending the Foundations of AI Literacy	International Conference on Information Systems (ICIS), A-Ranking (VHB JOURQUAL-4) https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/
Paper VI (Management)	Theoretical Foundations	Heyder , T.; Passlack, N. & Posegga, O.	2023 (published)	Ethical Management of Human-AI Interaction: A Theory Development Review	The Journal of Strategic Information Systems (JSIS), A-Ranking (VHB JOURQUAL-4) https://doi.org/10.1016/j.jsis.2023.101772
Paper VII (Dilemmas)		Hammerschmidt , T.	2024 (published)	Navigating the Nexus of Ethical Standards and Moral Values	Ethics and Information Technology (EIT), 4.2 Impact Factor https://doi.org/10.1007/s10676-025-09826-5
Paper VIII (Inequalities)		Hammerschmidt , T.; Stolz, K. & Posegga, O.	2024 (published)	Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI	Behaviour and Information Technology (BIT), B-Ranking (VHB JOURQUAL-4) https://doi.org/10.1080/0144929X.2025.2500451
Paper IX (Transformation)	Empirical Evidence	Hammerschmidt , T.; Passlack, N. & Posegga, O.	2025 (working paper, unpublished)	How Organizations Ethically Respond to Decision Dilemmas in AI Deployment	Working Paper (unpublished) <i>[previously 2nd revision round at JSIS]</i>
Paper X (Behavior)		Stolz, K.; Hammerschmidt T. & Posegga, O	2024 (published) Best Paper Nominee	How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making	Hawaii International Conference on System Sciences (HICSS), B-Ranking (VHB JOURQUAL-4) https://aisel.aisnet.org/hicss-57/in/social_shopping/3/

Table 1: Overview of this dissertation's articles.

The fifth section discusses the theoretical (Section 5.1), practical (Section 5.2), and methodological implications (Section 5.3) of all dissertation articles and outlines directions for future research based on the dissertation's limitations (Section 5.4). The dissertation ends with a conclusion in Section 6.

2 Theoretical Foundations and Related Research

This dissertation studies the understanding and management of ethics in human-AI interaction. The articles included in the cumulative dissertation rely on literature examining different ethical views such as normative ethical theories, providing structured frameworks for analyzing and evaluating ethical actions and decisions within human-AI interaction (Section 2.1). Additionally, it relates to literature referring to the sociomaterial notions to examine the complexity of different forms of human engagement in different cultural contexts with several types of AI technologies (Section 2.2). By considering different ontological positions, this dissertation aims to account for the differences in moral agency dependent on AI's capabilities and the perceptions that this creates in people.

2.1 The Role of Ethical Theories in Shaping Our Understanding of Ethical Human-AI Interaction

The exploration of ethics within the context of human-AI interaction begins with a critical understanding of ethics itself. Ethics involves the ability to critically evaluate moral values and guide actions accordingly (Churchill, 1999). This transcends simple reflection on moral principles, encompassing both the intent behind decisions and the resulting consequences when determining whether a decision or action is ethically “right” or “wrong” (Hosmer, 1987). Ethics is often understood as the science of morality, with morality defined by prevailing notions of “good” and “bad” within specific cultural, societal, and temporal contexts (Bryson, 2018). Thus, one's moral values are shaped explicitly by cultural norms or implicitly by societal values and evolve over time in response to social, political, or legal changes (Horváth, 2022). Moral values, therefore, are dynamic, reflecting the belief systems of a given moral community, including influences from religious institutions and governmental bodies (Horváth, 2022). This evolving nature of morality underscores its fluidity as it adapts to historical, political, and contemporary shifts, resulting in a unique ethical identity for each society (Bryson, 2018). Bryson (2018) notes that these moral systems influence not only individual decisions but also collective behaviors, which become codified in social norms. Consequently, the judgment of whether an action or decision is morally right or wrong depends on the specific ethical framework applied within a moral community (Hammerschmidt, Hafner, et al., 2025). The range of ethical frameworks—often classified as “ethical families,” “ethical theories,” or “ethical perspectives”—arises from different ontological positions that influence how we interpret morality (Hafner et al., 2024). Each perspective provides a distinct lens for evaluating ethical decisions and actions based on either ethical intention, behavior, or outcomes, whether examined in isolation or in relation to broader societal perspectives. As Schlagwein and Willcocks (2023, p. 233) stated,

Ethics offers various frameworks through which one can assess the rightness or wrongness of actions. Ethical theories can be classified into those that focus on the actions (means, processes, methods) as such and those that focus on the results (ends, goals, outcomes) achieved via these actions as primary grounds for ethical assessments.

Thus, the field of ethics encompasses various perspectives on evaluating moral conduct that help to examine different understandings of ethics in the context of human-AI interaction (Brendel et al., 2021).

Different approaches to ethics in human-AI interaction: There are various approaches to investigate ethics in different contexts of human-AI interaction that are somehow related to each other within the broader field of moral philosophy. While this dissertation primarily focuses on normative ethics, it also addresses meta-ethics, descriptive ethics, and applied ethics, as introduced here.

Meta-ethics investigates the foundational nature, scope, and meaning of ethics and morality, exploring the origin and meaning of moral values and ethical principles (e.g., “What does ‘good’ mean?”) and whether they are objective or subjective (Atkinson et al., 1969; Mackie, 1990). Rather than determining whether actions are right or wrong, meta-ethics examines the philosophical foundations of morality, exploring the origin and validity of ethical norms and moral values, thereby addressing the essence of morality itself (Atkinson et al., 1969; Tajalli, 2021), which is closely related to different ontological positions on how one sees the world. A notable figure in meta-ethics is G. E. Moore (1873–1958), whose work in *Principia Ethica* (1903) introduced the concept of the “naturalistic fallacy.” Moore argued that “goodness is a non-natural property,” meaning that it cannot be defined by naturalistic or empirical terms but rather exists as a simple, indefinable quality (Frankena, 1939, p. 473). This perspective aligns closely with substantial ontology, as Moore contends that goodness is an intrinsic, inherent property that exists independently of the physical world or natural relations (Frankena, 1939). Thus, from a substantialist position, moral properties such as goodness are seen as fundamental realities that exist autonomously, without being contingent upon relationships or external conditions (Hammerschmidt, Hafner, et al., 2025). Had Moore adhered to a relational ontology, he would have argued that goodness derives its meaning from its relations to other entities or properties such as its impact on human well-being or its connection to pleasure and other natural phenomena. This distinction is foundational to Cunneen’s (2019) argument that the ontological analysis must precede ethical inquiry, as it shapes how ethical concepts are understood and applied.

In the context of human-AI interaction, meta-ethical questions that can be answered from both ontological positions can arise, for instance, when considering moral relativism (Hammerschmidt, Hafner, et al., 2025). Moral relativists argue that incorporating multiple moral frameworks into AI law and regulation introduces the risk of ethical inconsistency, allowing people to justify any action based on a selective moral perspective (Conitzer et al., 2017; Serafimova, 2020). From a relational ontological

perspective, this relativism could be viewed as inherent to the interdependent nature of morality—moral values emerge and shift based on the interactions between different agents, contexts, and perspectives (Hammerschmidt, 2025). Conversely, from a substantial standpoint, the relativist critique underscores the potential risk of detaching AI from fixed moral principles that are viewed as universal and objective, independent of human-AI interactions (Hammerschmidt, 2025). Other meta-ethical questions emerge regarding whether AI can develop the capacity to grasp moral values and create its own moral framework based on its self-learning abilities or whether it merely simulates moral reasoning by reflecting the ethical preferences embedded in its design (Cook, 2023; Floridi & Sanders, 2004; Wehrens et al., 2023). This discussion intersects with the concept of artificial moral agency, questioning whether AI can be considered a moral agent or if it merely mimics ethical behavior (Allen et al., 2000). This discussion naturally connects to different ontological positions, as it raises questions about the nature of the existence of morality (substantial ontology) or its creation through interconnections (relational ontology) (Cunneen et al., 2019; Kieslich et al., 2022). A relational ontological perspective suggests that the ethical behavior of humans and AI technologies is shaped by their interactions (Heyder et al., 2023). Here, moral agency is dynamic and contextual, shaped by the evolving relationships and engagements between AI and human actors (Hammerschmidt, Hafner, et al., 2025). In contrast, a substantialist view treats AI and humans as distinct entities—humans possessing inherent virtues and AI governed by fixed ethical principles—that remain unaffected by their interactions (Heyder et al., 2023). AI, in this view, cannot possess moral agency because it lacks the intrinsic moral qualities that characterize human ethical reasoning (Hammerschmidt, Hafner, et al., 2025). As a result, the meta-ethical discourse surrounding AI intersects with broader ontological questions about the nature of morality. Whether one adopts a relational or substantial ontology profoundly influences how one understands ethical interactions between humans and AI (Cunneen et al., 2019). Distinct ontological frameworks lead to different interpretations of the “realities” of ethics within human-AI interactions, influencing how ethical behavior is defined (normative ethics), understood (descriptive ethics), and applied (applied ethics) in these contexts. Hence, this dissertation considers these meta-ethical questions when explaining ethical human-AI interaction.

Descriptive ethics, in contrast, takes an empirical approach, studying the moral beliefs, practices, and behaviors of individuals and societies (Hammerschmidt, Hafner, et al., 2025). It seeks to understand how different cultures and communities perceive moral issues and how these perceptions influence their actual behaviors and actions (Gert, 2005). Unlike normative ethics, descriptive ethics does not make evaluative judgments about moral beliefs but aims to describe and explain them (Weaver & Trevino, 1994). For instance, C. Gilligan, a feminist ethicist, is known for her work on the ethics of care, integrating diverse stakeholders’ perspectives into the empirical analysis of decision-making, ethical response, and moral learning while challenging dominant conceptions of morality such as Kantian or utilitarian theories (Gilligan, 2014). In the context of human-AI interaction, descriptive ethics helps to understand how humans engage with AI in specific contexts, examining ethical responses across

different cultures and societies regarding AI behavior (Constantinescu et al., 2021; Tóth et al., 2022). Typical topics of descriptive ethics involve responsibility attribution and perceptions of humans regarding AI in decision-making by empirically investigating moral expectations and behavior patterns related to AI technologies (Coeckelbergh, 2020; Forsyth et al., 2021). This dissertation especially links to descriptive ethics through those articles that focus on the empirical investigation of ethics within human-AI interactions.

In comparison, *applied ethics* bridges theoretical ethical principles with practical, real-world dilemmas (Singer, 1968, 2005). It entails the practical application of moral beliefs and ethical principles to specific domains such as medicine, business, and environmental issues, providing concrete solutions to ethical concerns or addressing specific moral dilemmas based on the frameworks provided by normative ethics (Singer, 1968, 2005). P. Singer is a well-known ethicist of applied ethics, recognized for his work on effective altruism, which focuses on how individuals can maximize their positive impact on others' well-being by using their resources (e.g., money, time) most effectively. This approach is linked to the utilitarian tradition (Singer, 1968) and to a relational ontology due to its focus on interdependencies (Hammerschmidt, 2025). In the context of human-AI interaction, applied ethics is reflected in current regulations such as the European AI Act, aimed at increasing transparency and fairness (Colmenarejo et al., 2022; Polyviou & Zamani, 2022), as well as in design guidelines for responsible and trustworthy AI (Buruk et al., 2020). Thus, applied ethics concerns regulation, design, or management, making human-AI interaction ethical from a relational ontological position that is closely related to this dissertation's second research question.

To summarize, *meta-ethics* addresses fundamental questions about the nature of ethics and morality, which are closely linked to different ontological positions that shape normative judgments about right and wrong (Hammerschmidt, Hafner, et al., 2025). *Applied ethics* focuses on the practical application of ethical principles to AI design and regulation, often drawing from deontological or consequentialist theories (Hammerschmidt, Hafner, et al., 2025). *Descriptive ethics* empirically explores moral values, beliefs, and behavior of individuals, frequently associated with virtue ethics, as well as how organizations ethically manage AI, typically aligning with consequentialist thought (Hammerschmidt, Hafner, et al., 2025). These approaches, in turn, often build on the inquiries of *normative ethics*. For example, addressing issues of bias and discriminatory AI outcomes requires an understanding of existing AI regulation and the application of deontological principles in AI design, which falls within the domain of *applied ethics*. *Descriptive ethics* is essential for examining how misconceptions about AI capabilities can lead humans to misjudge good and bad behavior in their interactions with AI, an issue related to virtue ethics that potentially results in the misuse of AI systems. Additionally, understanding the dynamics between human and AI agency, along with the increased integration of moral considerations within AI systems, necessitates engagement with theoretical and conceptual questions rooted in *meta-ethics*. Thus, normative ethical theories are interconnected with other ethical frameworks, whereby

meta-ethics provides ideas for the ontological analysis that precedes the analysis of normative ethical theories (Hammerschmidt, 2025). Consequently, this dissertation focuses on normative ethical theories for analyzing the ethical management of human-AI interaction, as these theories provide structured frameworks—such as deontology, consequentialism, and virtue ethics—that offer guidance for assessing what constitutes morally right or wrong actions and behaviors in specific situations.

Ethical theories shaping our understanding of moral conduct: The emphasis on normative ethical theories in the ethical management of human-AI interaction stems from their ability to offer clear and distinct frameworks for determining right and wrong actions from different perspectives. This is crucial for understanding and managing moral dilemmas that arise when conflicting moral values and ethical principles come into play (Brendel et al., 2021; Hammerschmidt, Hafner, et al., 2025; Mirbabaie et al., 2022). Normative ethics focuses on establishing frameworks for determining and assessing whether actions or decisions are morally right or wrong, depending on the outcomes of actions or the intention behind them (Behdadi & Munthe, 2020; Brendel et al., 2021). These frameworks offer structured perspectives for analyzing and explaining differences in moral values, societal norms, and given principles, ultimately informing ethical decision-making and determining the appropriate course of action in diverse scenarios (Brendel et al., 2021; Kagan, 1992). Thereby, it investigates how one ought to act and what moral standards should guide behavior. The interplay between virtue ethics, deontology, and consequentialism—the three most common normative ethical theories—demonstrates how ethical analysis can integrate considerations of duty (deontology), outcomes (consequentialism), and character (virtue ethics) (Bankins & Formosa, 2023; Kagan, 1992, 2018).

Virtue ethics addresses questions such as “What kind of person I want to be?” by emphasizing the role of individual character and inherent ethical motivations that guide behavior and actions (Darwall, 2005). Aristotle (384–322 BCE) developed this ethical theory, highlighting the cultivation of virtuous character traits such as courage and honesty, as central to moral behavior (Aristoteles et al., 1962; Simpson, 1992). These virtues can evolve over time into certain norms and guidelines, intersecting with deontological ethics when they form duties (Eitel-Porter, 2021). While AI systems cannot possess inherent virtue ethics (Constantinescu et al., 2021), the self-learning capabilities of AI—drawing on human input and feedback—may allow for a collective form of virtue ethics to be integrated into AI (Constantinescu et al., 2021; Etzioni & Etzioni, 2017). In the context of human-AI interactions, virtue ethics centers on the moral intentions of individuals in designing and using AI technologies, influencing ethical or nonethical behavior of individuals (Coeckelbergh, 2020; Hagendorff, 2020). Thus, virtue ethics plays a significant role in guiding the ethical design and responsible use of AI technologies (Hammerschmidt, Hafner, et al., 2025). However, as virtue ethics is deeply rooted in personal motivations and subjective moral judgments, it can be challenging to quantify or measure its impact on AI outcomes, making it difficult to capture in empirical analyses—especially descriptive ethics.

In comparison, *deontology* posits that actions are inherently right or wrong, with morally acceptable actions being those that can be universalized without contradictions, thus forming a “universal law,” as outlined in Kant’s (1991) categorical imperative (Gaus, 2001; Unver, 2023). Deontology is grounded in duty and obligation, judging the morality of actions based on universal ethical principles that can become general law (Gaus, 2001; Unver, 2023). Regulations, such as the European AI Act, can be seen as a form of deontological ethics within the context of human-AI interaction, as they establish legal frameworks that define permissible actions and are then interpreted in specific cases (Koniakou, 2023). As AI capabilities develop, there is an ongoing debate over whether and how Kant’s categorical imperative can be applied to AI. Some argue AI should adhere strictly to predefined ethical guidelines, while others suggest leveraging AI’s learning capabilities for ethical reasoning by design (Dyrkolbotn et al., 2018).

Consequentialism, by contrast, determines the morality of actions based on their outcomes or consequences (Mill et al., 1989). Various forms of consequentialism differ in what they consider the relevant consequences for judging moral actions (Höffe, 2013). Utilitarianism, the best-known variant of consequentialism, focuses on maximizing overall happiness and pleasure for the greatest number of people, a concept known as Mill’s (1989) “greatest happiness principle” (Bentham, 1988; Höffe, 2013). In the context of human-AI interaction, building a “utilitarian” AI would entail reducing morality to a cost-benefit analysis, raising questions about whether happiness can be unequivocally measured as a matter of goodness (Serafimova, 2020). However, as Hauer (2022) argues, utilitarianism can offer valuable guidance in contexts such as autonomous driving, where other ethical justifications may fail to provide clarity.

In summary, understanding the interplay of these normative ethical theories in the context of human-AI interaction is critical, as AI decision-making requires careful evaluation of both the processes involved and their societal consequences (Rodgers et al., 2023). An overemphasis on virtue ethics may prioritize the moral intentions of AI designers while overlooking diverse values from other moral communities, leading to moral relativism (Conitzer et al., 2017; Hammerschmidt, Hafner, et al., 2025; Serafimova, 2020). Similarly, deontological approaches may offer clear ethical guidelines for AI but risk oversimplifying complex moral dilemmas (Millar, 2016; Swanepoel, 2021). Conversely, consequentialisms focus on maximizing benefits for the greatest number of stakeholders but may reduce morality to a cost-benefit analysis (Serafimova, 2020) or restrict individual moral autonomy by emphasizing collective norms (Terzis, 2020). Therefore, navigating the ethics of human-AI interaction requires balancing these competing normative perspectives to manage moral dilemmas and ethical challenges effectively.

Investigating moral dilemmas through normative ethical theories: Moral dilemmas emerge when individuals or organizations must choose between conflicting moral values or ethical standards, each of which may result in undesirable outcomes (Hammerschmidt, Passlack, et al., submitted; Kegel &

Ghanem, 2024; Z. Zhang et al., 2022). In the context of human-AI interaction, these dilemmas are especially intricate, as the unpredictable side effects of AI technologies often force decision-makers to accept moral trade-offs (Kegel & Ghanem, 2024; Z. Zhang et al., 2022). Moral trade-offs occur when individuals or organizations must choose between conflicting ethical principles or values, knowing that fulfilling one moral obligation may lead to the violation of another (Hammerschmidt, 2025; Hammerschmidt, Passlack, et al., submitted). For example, using AI in hiring processes might increase efficiency and reduce bias on a broad scale, but it could also unintentionally marginalize certain groups if the algorithms fail to account for nuanced social factors. In such scenarios, decision-makers face a trade-off between achieving greater fairness overall and ensuring justice for all individuals (Hammerschmidt, 2025; Hammerschmidt, Passlack, et al., submitted). Hence, these dilemmas arise when decision-makers must choose between undesirable options, each carrying unavoidable negative consequences for the organization or its stakeholders (Bonnefon et al., 2024; Marabelli & Galliers, 2017; Shank et al., 2019). The growing integration of AI into decision-making processes further adds complexity since, while AI lacks traditional moral agency, it can significantly influence human choices and actions (Cook, 2023; Susser, 2019). Researchers have highlighted the importance of aligning various normative ethical approaches to mitigate these dilemmas and prevent responsibility gaps (Buhmann & Fieseler, 2023; Søvik, 2022). Without alignment, ethical standards may lack social acceptance and give rise to value conflicts (Buruk et al., 2020; Córdova & Vicari, 2022; Z. Zhang et al., 2022). These conflicts occur when ethical guidelines, while intended to guide responsible behavior, clash with individual or societal moral values (McDonald & Pan, 2020; Shank & Gott, 2020). For example, an ethical framework in one context might justify certain actions that, in another context, are considered morally wrong, creating dilemmas about the most ethical course of action (McDonald & Pan, 2020; Shank & Gott, 2020). Ethical standards (linked to deontology) serve as guidelines for evaluating responsible behavior, but they are not immune to conflict, especially when assigning moral blame for wrongdoing (Hammerschmidt, 2025; McDonald & Pan, 2020).

Some scholars argue that seeking perfect alignment between ethical standards and moral values imposes unnecessary constraints, establishing a baseline for acceptable behavior that may fall short of higher individual moral ideals (Constantinescu et al., 2021; Meijer et al., 2023). While alignment can help reduce conflicts, it is not a universal solution (Hammerschmidt, 2025). Moral dilemmas frequently occur when AI is misused or employed for unethical purposes (Kegel & Ghanem, 2024). For instance, AI technologies may be exploited to manipulate social media algorithms, spread misinformation, or promote harmful ideologies (Wilson & Land, 2021). In such cases, the mere presence of moral agency is insufficient to prevent unethical behavior (Hammerschmidt, 2025). The ethical standards guiding both human and AI decision-making become crucial to mitigating these risks and ensuring the responsible use of AI technologies (Meijer et al., 2023; Peters et al., 2020). This can be further applied to ambidextrous decision-making, where deciders have to balance exploitation strategies that prioritize the use of AI for optimization but may miss AI's potential to generate benefits for a greater number of

organizational stakeholders with exploration strategies that focus on AI's deployment for innovation but may involve ethical risks (Hammerschmidt, Passlack, et al., submitted). Additionally, AI systems can exacerbate moral dilemmas by amplifying existing biases in their algorithms (Zheng et al., 2024). An example is the use of AI in crime detection, which, while fulfilling the moral imperative of enhancing public safety, can lead to discriminatory practices against marginalized communities or infringe on individual privacy rights (B. Zhang et al., 2021). These dilemmas require careful ethical consideration, as AI may unintentionally cause harm even when it aims to promote morally justified outcomes (Giermindl et al., 2022; Mayer et al., 2020). Therefore, considering the interplay between normative ethical theories is essential in navigating moral dilemmas, as together, these theories provide a comprehensive framework for addressing the multifaceted moral dilemmas in human-AI interaction (Hammerschmidt, 2025).

Ethical capabilities for realistic perceptions of AI's moral agency: As AI capabilities continue to develop and surpass the capabilities of previous systems (White et al., 2023), their abilities to imitate human behavior and values based on training data (Chedrawi & Haddad, 2022) have intensified debates about the distinction between AI and human intelligence (Kaplan & Haenlein, 2019). While AI can be designed to perform complex cognitive tasks (Rai et al., 2019), it lacks the multidimensional intelligence, intuition, and emotional depth that characterize human nature (Gardner, 2011). However, many people develop unrealistic perceptions and misconceptions about AI's capabilities (Hammerschmidt et al., 2024), including its potential for moral reasoning and decision-making (Diederich et al., 2022; Willcocks, 2020). This misunderstanding extends to ethical considerations, as individuals often attribute moral agency and responsibility to AI when it is involved in decisions with ethical implications (Giroux et al., 2022; J. S. Roberts & Montoya, 2022). Hauer (2022, p. 1) points out that "if a task requires some form of moral authority when it is performed by humans, its full automation, transferring the same task to autonomous machines, platforms, and AI algorithms, necessarily implies the transfer of moral competence." This perception complicates the human-AI relationship, as people may start to view AI as capable of making ethical judgments (Hammerschmidt et al., 2024), blurring the lines between human and machine moral agency (Coeckelbergh, 2020; Søvik, 2022). Further, such a lack of knowledge and inadequate skills among individuals to deal with AI and assess its capabilities can create gaps between those who appropriately and effectively use AI technologies and those who cannot, illustrating inequalities in benefiting equally from AI progress (Hammerschmidt, Stolz et al., 2025). To navigate this complex landscape, it is essential for individuals to cultivate ethical competencies that allow for a critical evaluation of AI technologies (Heyder & Posegga, 2021; Ng et al., 2023). Thus, people need to develop new competencies that enable them to critically evaluate AI technologies and realistically assess AI's capabilities.

These capabilities are closely linked to the concept of AI literacy concerning competencies, knowledge, and skills, making individuals able to work effectively with AI technologies and evaluate

their impact (Long & Magerko, 2020; Pinski et al., 2024). As part of AI literacy, ethical competencies concern the ability to “identify and describe different perspectives on the key ethical issues surrounding AI (i.e., privacy, employment, misinformation, the singularity, ethical decision-making, diversity, bias, transparency, accountability)” (Long & Magerko, 2020, p. 7). By gaining knowledge of these various ethical concerns and the ability to distinguish between the unique differences in AI and human moral agency, individuals can more accurately and realistically assess AI capabilities (Hammerschmidt et al., 2024). This further helps to establish knowledge of different ethical perspectives guiding human intentions and AI outcomes and to distinguish between ethical and unethical uses of AI.

In conclusion, various ethical approaches exist that can guide ethical analyses for understanding and shaping human-AI interactions by offering structured perspectives on moral conduct. Meta-ethics addresses foundational questions of morality and the understanding of ethics based on different ontological positions, applied ethics bridges theory with real-world challenges, and descriptive ethics empirically investigates moral values and behavior. These approaches are frequently interwoven with normative ethical theories such as deontology, consequentialism, and virtue ethics, which help navigate moral dilemmas by evaluating actions based on duties, outcomes, and character. Ultimately, knowledge of the interplay of these ethical perspectives is vital for addressing complex ethical challenges in AI decision-making, managing potential moral dilemmas of unintended negative impacts, and fostering realistic perceptions of AI’s moral agency compared to that of humans.

Table 2 provides an overview of how the dissertation’s articles refer to the different ethical perspectives:

Perspective	Paper	Description
Descriptive Ethics	III, IV, V, IX, and X (implicitly)	• Paper III examines individual perceptions of AI technologies and highlights the importance of specific ethics-related capabilities in fostering more realistic narratives about AI’s potential and increasing awareness of biases in AI systems. • Paper IV investigates how leadership decision-making and ethical capacities influence an organization’s ability to adapt to AI transformations, underscoring the significance of intangible ethical resources in navigating AI implementation. • Paper V emphasizes the crucial role of critically assessing AI’s outputs and implications in enhancing employees’ AI literacy, linking this to their ability to recognize ethical issues. • Paper IX explores how different stakeholders approach decision-making dilemmas during AI deployment and utilization, shaped by the normative ethical frameworks they predominantly follow. • Paper X demonstrates how nudging mechanisms embedded in conversational agents influence decision-making toward more sustainable product choices, showing how AI design can affect users’ ethical behavior.
Applied Ethics	VIII and IX (implicitly)	• Paper VIII fits within the applied ethics category by discussing the reduction of inequalities that impede sustainable outcomes in human-AI interactions, particularly in addressing the social, ethical, and environmental impacts of AI in alignment with the UN’s Sustainable Development Goals (SDGs).

		• Paper IX can also be linked to applied ethics, as it examines managerial strategies for addressing moral dilemmas arising from the evolving nature of human-AI interaction in organizational settings.
Normative Ethics (specifically deontology, virtue ethics, and utilitarianism)	I, II, VI, and VII (explicitly)	• Paper I explicitly engages with four normative ethical theories—deontology, virtue ethics, consequentialism, and contractarianism—to illustrate how varying ethical perspectives can lead to moral uncertainties in AI usage. • Paper II focuses on deontology, virtue ethics, and consequentialism to explore the diverse interpretations of AI ethics among researchers and how these perspectives shape ethical evaluations of AI. • Paper VI analyzes the interplay between deontology and virtue ethics within the context of human-AI interactions, providing insights into how both ethical theories can coexist and inform ethical AI management. • Paper VII uses deontology, virtue ethics, and consequentialism to demonstrate how different moral maxims based on these theories can lead to moral dilemmas in AI applications.
	III, IV, V, VIII, and X (implicitly)	• Paper III, though not explicitly addressing virtue ethics, aligns closely with it by examining how individuals' perceptions of AI, shaped by ethical narratives, are influenced by their inherent beliefs about AI's moral dimensions. • Paper V implicitly connects to virtue ethics, as it highlights individuals' capacities for ethical reasoning, understanding AI's impact, and recognizing ethical issues as key components of AI literacy. • Paper IV also implicitly connects to virtue ethics, emphasizing that intangible resources such as organizational ethics and culture are essential for managing AI transformations. • Paper VIII, while not directly discussing normative ethical theories, implicitly relates to deontology through its focus on AI governance to reduce inequalities, to virtue ethics at the individual level, and to consequentialism at the organizational level, particularly in addressing market dynamics. • Paper X touches upon all three normative ethical theories by examining the ethical design of conversational agents using nudges to promote sustainable consumption. This involves utilitarian goals (promoting consumption), deontological design principles, and customer moral beliefs tied to virtue ethics.
Meta-ethics	V and VII (implicitly)	• Paper V explains the interaction between deontology and virtue ethics in human-AI interactions, implicitly linking them to ontological considerations from meta-ethics. • Paper VII implicitly connects meta-ethics and normative ethical theories by exploring how differing ontological and normative views lead to moral dilemmas, particularly in relation to AI's agency.
All Ethical Perspectives	I and II (explicitly)	• Paper I bridges meta-, applied, descriptive, and normative ethics, providing a comprehensive overview of how different ethical perspectives contribute to moral uncertainties in human-AI interactions. • Paper II highlights the importance of examining AI ethics through applied, descriptive, and meta-ethical perspectives, reflecting the interdependence between theoretical justification and empirical interpretation in AI ethics.

Table 2: Overview of how the dissertation's articles refer to different ethical perspectives.

2.2 The Sociomaterial Notions Explaining Ethics and Moral Agency in Human-AI Interactions

This dissertation builds upon and extends the theoretical framework developed by Heyder et al. (2023), integrating the notions of sociomateriality into sociotechnical systems theory to create a framework for analyzing the understanding and management of ethical human-AI interactions. This approach draws on Leonardi's (2012) work, melding sociotechnical systems theory with sociomateriality discourse, and integrates Jones' (2014) perspective on the five weak and strong notions of sociomateriality: inseparability, relationality, performativity, materiality, and practices. Both

concepts are grounded in the discourse on sociomateriality advanced by W. Orlikowski and S. Scott (e.g., Orlikowski, 2007, 2010; Orlikowski & Scott, 2008). While sociotechnical systems theory helps to examine organizations as complex systems comprising both *social subsystems* (e.g., communication networks, organizational structures, market dynamics) and technical subsystems incorporating *sociomaterial practices* of social and technical actors, sociomateriality provides a deeper understanding of how technologies as *material entities* and humans as *social entities* co-construct sociomaterial practices through their entanglement, informed by social phenomena such as culture, meanings, and beliefs (Leonardi, 2012). These sociomaterial practices, in turn, influence the social subsystem and vice versa (Leonardi, 2012). The social subsystem, thereby, is a nested social system comprising several micro- and macro-systems (e.g., work system, organizational system, societal system), impacting sociomaterial practices through various social phenomena of human-related objects (e.g., cultural beliefs, managerial decisions) and nonhuman objects (e.g., laws, codes of conduct, norms, guidelines) at multiple levels (Winter et al., 2014). Thus, within the social subsystem, there is a system level (including, for instance, the work, organizational, and societal systems) and an object level (with human-related and nonhuman objects). Thereby, diverse moral values, ethical principles, cultural norms, organizational structures, and strategies, shaped by various organizational stakeholders within these nested systems, influence sociomaterial practices as “institutionalized ideas about how people could and should relate to one another” (Leonardi, 2012, p. 40). Cecez-Kecmanovic (2014, p. 814) argues, “The sociotechnical systems approach is important both as a precursor to sociomaterial thinking and also as a foundation for a particular view on sociomateriality.” By relating material elements with social phenomena, this framework enables the analysis of how AI technologies and humans produce ethical or unethical outcomes and how these interactions can be managed (un)ethically in organizational settings (Heyder et al., 2023). The focus thus lies on the relationship between how these elements produce distinct realities, including ethical choices shaped by the specific worlds created through their entanglement (Barad, 2007).

In this dissertation, an organization is viewed as a sociotechnical system, while interactions between AI technologies and human users are understood through sociomaterial entanglement, melding the materiality of technologies’ properties “with institutions norms, discourses, and all other phenomena we typically define as ‘social’” (Leonardi, 2012, p. 30). Using this lens helps to identify interdependencies among humans (whose actions may be informed by virtue ethics based on inherent moral values), technologies (embodying deontological principles through ethical design), and various organizational factors, such as AI strategies (which may be driven by utilitarian goals), corporate culture (grounded in virtue ethics), and certain guidelines for workforce AI usage (rooted in deontology), and illustrates how this creates different ontologies (Callon, 1999; Cecez-Kecmanovic et al., 2014). Thereby, the remaining part of Section 2.2 starts with a description of sociomaterial agency and the metaphor of imbrication, followed by discussions on AI versus human moral agency and the non-neutral co-constitution of AI’s role in shaping ethical outcomes. This section then outlines the background on two different ontological

positions shaping our understanding of human and AI moral agency (related to relational and substantial ontology) and demonstrates how the sociomaterial notions can be used to provide insights into the ethical management of human-AI interactions based on these two positions.

Social and material agency and the metaphor of imbrication: This dissertation centers on how the entanglement of social (human users) and material (AI technologies) entities produces ethical behavior and outcomes within sociomaterial practices, influenced by the distinct agency of each entity and how this is further shaped by the organizational environment, understood as a social subsystem (Leonardi, 2012). Leonardi (2011, p. 151) introduces the “metaphor of imbrication” to describe the different ways in which the entanglement of human and material entities produces outcomes within sociomaterial practices based on their distinct agencies. Besides the different capabilities of humans and technologies—specifically the intentionality of social entities forming human agency—the materiality of the technological artifact creates the material agency (Leonardi, 2011, 2012). Thus, at the core of sociomaterial entanglement lies the concept of agency, illustrating “what’s going on inside” an agent, which refers to an entity’s intentions, reasons, and goals that shape its behaviors, actions, and decisions in relation to the interaction (Gogoshin, 2021, p. 2). Human agency, as the agency of social entities, is linked to the concept of inner free will, which is influenced by ethical ideologies, intentions, and moral values that impact how individuals respond to technologies and their outcomes, allowing them to choose whether they behave ethically or not. These elements are central to virtue ethics (Coeckelbergh, 2020; Eitel-Porter, 2021; Flathmann et al., 2021). This is what Leonardi (2011, 2012) describes as the intentionality of people guiding human agency. In contrast, material agency, as exhibited by AI technologies as material entities, refers to “the way objects act when humans provoke it” (Leonardi, 2012, p. 37), which can encompass ethical actions rooted in deontological principles but lacks intentionality (Heyder et al., 2023). This is what Leonardi (2011, 2012) describes as materiality, referring to the properties of the technological artifact. The interaction between these forms of agency lies at the heart of sociomaterial entanglement, highlighting how human intentions and technological materiality co-create distinct behaviors and outcomes.

AI versus human moral agency: The idea of relating AI’s agency to ethics and morality goes back to the discourse surrounding Moor’s differentiation between “implicit” and “explicit ethical agents” (2006, pp. 19–20). According to this view, AI can act as an explicit ethical agent that operates on behalf of ethical principles defined by humans to make “judgments of duty” (aligning with the concept of AI as a moral patient) or as an implicit ethical agent that “act[s] ethically because its internal functions implicitly promote ethical behavior—or at least avoid unethical behavior [as] it has, to a limited extent, virtues” through learning from its human users’ data input and feedback (aligning with the idea of AI as a moral agent) (Moor, 2006, pp. 19–20). Moral agency refers to an entity’s ability to follow moral maxims, make ethical judgments, and take responsibility for its actions, shaped by its cultural background and the societal values within the “moral community” in which the entity is embedded,

forming its belief system with perceptions of “good” and “bad” (Behdadi & Munthe, 2020; Champagne & Tonkens, 2023, p. 27:8; Mabaso, 2021). The term “moral agency” is often synonymously used with ethical agency, which concerns reflecting on what is “right” and “wrong” by weighing different moral maxims against the often invisible background of one’s belief system (Bankins & Formosa, 2023; Cheruvalath, 2023; Dyrkolbotn et al., 2018). Thus, ethical agency is about reflecting on different moral agencies, which can be interpreted through various ethical theories (e.g., Bankins & Formosa, 2023; Bilal et al., 2021; Cheruvalath, 2023), as introduced in Section 2.1. While moral agency is traditionally ascribed to humans as social entities, AI technologies, as material entities, act on ethical principles that are based on the ethical perceptions of developers and regulators embedded in the design of the AI system. This also illustrates some moral boundaries, a phenomenon referred to as artificial moral agency (Behdadi & Munthe, 2020; Floridi & Sanders, 2004). Additionally, through AI’s ability to learn unsupervised from human data input and feedback, it can further refer to human moral values in their responses (Etzioni & Etzioni, 2017; Moor, 2006). Conflicts may arise when moral values embedded in user data from different moral communities, incorporating different belief systems of what is morally good and bad, clash together—an issue commonly explored through the trolley dilemma in the context of autonomous driving (Cunneen et al., 2019). This raises questions about whether AI systems should be designed to recognize and prioritize one belief system of a moral community over another (Hammerschmidt, Hafner, et al., 2025). Even though AI can process data containing moral considerations from various human users and generate outcomes that might appear to reflect ethical reasoning (Dyrkolbotn et al., 2018), its moral agency fundamentally differs from that of humans (Hammerschmidt, 2025). AI lacks the capacity to act “freely (as in ‘having free will’) or ‘know’ (as in ‘being aware of’) what it is doing” (Coeckelbergh, 2020, p. 2055). As a result, AI cannot comprehend or engage with moral values in the same way that humans do, since “a computationally correct answer is not a moral outcome” (Puri, 2020, p. 52).

The non-neutral co-constitution of AI agency: The growing (semi-autonomous) capabilities of AI technologies, which enable the systems to perform tasks with limited human oversight, challenge the traditional view of technological artifacts as neutral entities that only respond to human interaction (Heyder et al., 2023). Cecez-Kecmanovic (2014) described this shift as the non-neutral co-constitution of material agency. Unlike traditional technologies, AI systems are not merely neutral instruments shaped by human intent (as “moral patients”) but rather active participants in shaping social and ethical outcomes as “moral agents,” introducing moral considerations that are not fully predictable or controllable by humans (Allen et al., 2000). Thus, non-neutrality here refers to the idea that AI technologies, compared to previous technologies, actively participate in shaping ethical behavior and implications, while co-constitution implies that the ethical behaviors and consequences that arise from human-AI interaction are mutually shaped by both human and technological entities. One example of this non-neutral co-constitution of AI technologies can be found in generative AI models such as OpenAI’s generative pretrained transformer (GPT), when compared to Orlikowski’s (2007) discussion

of Google's PageRank algorithm. Orlikowski (2007) illustrates how Google's PageRank algorithm (a material entity) continually updates to reflect the choices made by users (a social entity) as part of the social process. This process exemplifies the co-constitution of technology and society: the user's search behavior influences the algorithm's future outputs, and, in turn, the algorithm's outputs shape the user's behavior. In the case of generative AI models, this co-constitution is even more pronounced and additionally becomes non-neutral as generative AI models function in a more complex, autonomous, and opaque manner (Schlagwein & Willcocks, 2023). This means that although human input is essential for training AI systems and for predefining their ethical parameters, the internal processes that produce outputs are disconnected from direct human intervention and social processes, potentially creating black-box mechanisms that even developers may find inscrutable (Asatiani et al., 2020; Rai, 2020). Human involvement, thereby, typically occurs only when AI outputs require correction, illustrating that human agency remains in place to oversee and override the autonomous actions of AI. This contrasts with the dynamics of PageRank, where the algorithm's updates are clearly linked to user interactions as moral patients. As a result, AI technologies do not consistently act as moral patients; they do not always perform actions based on human ethical frameworks, thus extending the traditional boundaries of moral involvement (Allen et al., 2000; Behdadi & Munthe, 2020; Floridi & Sanders, 2004). Particularly when AI outcomes carry moral implications—whether positive or negative, such as ChatGPT's refusal to generate content that aligns with extremist ideologies or other harmful requests (Krügel et al., 2023; Stahl & Eke, 2024) or Microsoft's chatbot Tay exhibiting inappropriate behavior when launched on Twitter (now: X) (Shank & Gott, 2020; Suárez-Gonzalo et al., 2019)—people tend to attribute moral responsibility to AI technologies, perceiving them as moral agents (Allen et al., 2000; Behdadi & Munthe, 2020; Coeckelbergh, 2020; Floridi & Sanders, 2004). In both cases, the AI models would not produce such outputs without human-defined ethical parameters (as seen with ChatGPT) or human intervention to correct ethical misconduct (as was necessary with Tay). These examples illustrate that human intervention is necessary because AI cannot embody the concept of morality on its own. Nevertheless, users may perceive these systems as moral agents capable of exhibiting moral behavior—whether right or wrong—despite their reliance on human oversight and correction. This illustrates a non-neutral co-constitution between technologies and social actors. With AI's increasing agency—in terms of autonomy, responsibility, and knowledge—the concept of “constitutive entanglement” introduced by Orlikowski (2007, p. 1435) does not imply equality between humans and technologies (Cecez-Kecmanovic et al., 2014). This shift challenges previous assumptions about agency and blurs the lines of responsibility concerning the development, management, and ethical use of AI technologies. It raises critical questions about who should be held accountable for unintended, potentially negative, outcomes arising from human-AI interaction (Horneber & Laumer, 2023; McGrath & Whitty, 2018; Novelli et al., 2023). Thus, “[a]gencies are not attributes [of either humans or technologies] but ongoing reconfigurations of the world” (Barad, 2003, p. 818). This perspective emphasizes that the interplay

between human and technological agency shapes both ethical and nonethical behaviors and outcomes in everyday human-technology interactions (Cecez-Kecmanovic et al., 2014).

Morality and agency as ontological matters: Ontological frameworks play a crucial role in defining the agency and morality of both humans and AI. Different cultural environments and technological artifacts give rise to varied realities on ethical AI design, deployment, and usage (Cunneen et al., 2019). Ontology, a branch of philosophy that examines the nature of being and existence, provides foundational ideas for understanding the relationships and properties that define the agency of material (AI) and social (human) entities within specific domains such as organizations as sociotechnical systems (Orlikowski, 2010). The organizational environment, comprising actors, objects, and artifacts, creates multiple realities that shape the ontological understanding of the ethical standards and moral values guiding both human and AI moral agency (Cunneen et al., 2019). As Latour and Venn (2002, p. 254) assert,

Morality is no more human than technology, in the sense that it would originate from an already constituted human who would be master of itself as well as of the universe. . . . Morality and technology are ontological categories . . . and the human comes out of these modes, it is not at their origin. Or rather, it cannot become human except on condition of opening itself to these ways of being which overflow it from all sides and to which it may choose to be attached – but then at the risk of losing its soul.

This perspective suggests that humans do not originate morality but become moral beings through their engagement with both technological and moral frameworks. Consequently, ontological positions influence perceptions of what constitutes agency and morality for humans and technologies and how their interplay leads to ethical or unethical behavior in human-AI interactions.

Combining, rather than comparing, ontological positions via sociomaterial notions: This dissertation focuses on integrating two main ontological positions—substantial ontology and relational ontology—to explain our understanding of and the management of ethical human-AI interactions. *Substantial ontology*, associated with *weak sociomateriality*, posits that humans and technologies are independent, discrete, and self-contained entities with intrinsic properties that remain stable (Cecez-Kecmanovic et al., 2014). In this position, the interaction between humans and AI serves only as moderator, influencing behavior without fundamentally altering either the ethical principles of AI technologies or the ethical ideologies of humans (Cecez-Kecmanovic et al., 2014; Fabri et al., 2023). Proponents of this position, such as critical realists Faulkner and Runde (2010) and Mutch (2010, 2013), argue that while social and material entities interact, they retain their distinct characteristics in separate existences. In contrast, *relational ontology*, aligned with *strong sociomateriality*, posits that social and material entities are inseparable and co-constituted through their constitutive entanglement (Cecez-Kecmanovic et al., 2014). This perspective is supported by agential realists like Barad (2003, 2007),

Orlikowski (2007, 2010), and Latour (2005), emphasizing the formation of sociomaterial assemblages in which humans and technologies mutually shape each other's inherent properties, including ethical orientations and roles. As Orlikowski (2007, p. 1437) states,

[T]he social and the material are constitutively entangled in everyday life. A position of constitutive entanglement does not privilege either humans or technology (in one-way interactions), nor does it link them through a form of mutual reciprocation (in two-way interactions). Instead, the social and the material are considered to be inextricably related—there is no social that is not also material, and no material that is not also social.

Given the aforementioned non-neutral co-constitution of ethics—where AI technologies with varying degrees of autonomy and self-learning abilities interact with humans from diverse cultural and moral contexts—it is crucial to integrate both ontological positions to understand how these interactions lead to differing views on human and AI agency and, consequently, on ethical human-AI interaction (Chedrawi & Haddad, 2022; Fabri et al., 2023). By combining, rather than comparing, these ontological frameworks, this dissertation prioritizes ontological analysis as a precursor to ethical evaluation in the context of AI technologies (Cunneen et al., 2019), offering distinct yet potentially equally valid insights (Cecez-Kecmanovic et al., 2014; M. Jones, 2014). Thereby, the *weak* and *strong* sociomaterial notions as outlined by Jones (2014) based on Orlikowski's work—such as inseparability, relationality, performativity, materiality, and practices—are employed to provide deeper insights into the complex relationships between the social and material dimensions of ethical human-AI interaction based on both ontological positions (Cecez-Kecmanovic et al., 2014; M. Jones, 2014). Integrating these sociomaterial notions with Leonardi's (2012) conceptualization of sociotechnical systems allows for a closer examination of the ethical interdependencies between the social subsystem of organizations (e.g., managers acting as ethical role models or guidelines for ethical AI usage) and the sociomaterial entanglement of humans interacting with AI technologies (Heyder et al., 2023).

Investigating ethics through the weak and strong sociomaterial notions: Starting with the sociomaterial notion of *inseparability*, which describes the multiple ways in which social and material entities are intertwined during interaction (M. Jones, 2014). This notion is closely related to the concept of imbrication, reflecting overlapping expressions of human and material agency within sociomaterial practices (Leonardi, 2012). Strong inseparability signifies a profound interdependence between the material and the social, wherein “[m]ateriality is thus viewed as integral to human activity, and there is no social action that does not entail material means” (M. Jones, 2014, p. 898). At its most intense, this results in what is termed a “sociomaterial assemblage” (Orlikowski & Scott, 2008, p. 462). In the context of human-AI interaction, Chedrawi and Haddad (2022) describe an intra-relationship between humans and AI, where AI's ability to recognize emotions and inspire thoughts significantly impacts social behaviors, demonstrating the increased power of AI agency over human agency. Fabri et al. (2023) refer

to this dynamic as human-AI hybrids, where human decision-making is so deeply interwoven with AI-generated insights that the boundaries between human and AI agency blur. For example, consider mental health chatbots used in therapeutic contexts, where virtual AI counselors, equipped with emotion-recognition algorithms, adapt their responses based on users' emotional or physical states (e.g., Grodniewicz & Hohol, 2023). The AI's ability to recognize, interpret, and respond to human emotions or health data in real time directly influences the therapeutic or medical outcomes, shaping decisions that have ethical implications for the patient's well-being (e.g., Grodniewicz & Hohol, 2023). As Grodniewicz and Hohol (2023, p. 7) stated,

Another problem worth examining is that, while mimicry might increase users' engagement at the beginning, it might also have a detrimental effect when things do not go well, and the users' situation is not improving despite their efforts. In such a case feeling as if somebody cares, might not be enough.

This leads to an ethical responsibility for patient care among those institutions providing and using such mental health chatbots in therapeutic contexts, as they must ensure that AI's response aligns with ethical norms, such as promoting the patient's well-being and avoiding harm, since neither the human nor the AI can be viewed as fully independent in shaping ethical outcomes in this scenario. In contrast, *weak inseparability* suggests that each entity operates more independently, with the imbrication being asynchronous, such as humans preparing AI by inputting and training data and then evaluating its performance after the AI has completed its tasks (Heyder et al., 2023). Chedrawi and Haddad (2022) describe this as an interrelationship where humans provide the necessary data to train the AI systems, which then perform cognitive and analytical tasks autonomously. Human and AI learning processes occur separately, maintaining distinct and independent operations (Fabri et al., 2023). Thus, human agency remains autonomous and more powerful than AI when looking at the total process, in which there is solely a sequential interaction (Fabri et al., 2023). An example of this weak inseparability might involve AI-powered content moderation systems on social media platforms. Human moderators establish ethical guidelines (based on moral values) that the AI follows when scanning content for harmful speech or misinformation (e.g., Wilson & Land, 2021; Zarnoufi et al., 2020). However, the AI operates autonomously based on predefined ethical parameters, and humans intervene only if the AI's actions need correction. In this case, while the AI's decisions are rooted in ethical principles provided by humans, the human agency remains dominant, as it ultimately oversees and can override the AI's autonomous actions. The interaction is sequential rather than simultaneous, demonstrating the weak inseparability of human moral frameworks and AI's ethical parameters (Fabri et al., 2023). Thus, the sociomaterial notion of inseparability can be used to analyze the multiple ways in which human users' ethical interaction behavior (based on their virtue ethics) is intertwined with AI's adherence to given ethical parameters and design principles (based on deontology).

Continuing with the sociomaterial notion of *relationality*, which builds on the concept of inseparability, this perspective focuses on the results and outcomes that emerge from the interaction between social and material entities (M. Jones, 2014). It examines how the entanglement between these entities affects their properties, capabilities, behaviors, norms, and characteristics (M. Jones, 2014). Scholars like Orlikowski and Scott (2008) have explored how these interactions lead to changes in the properties of either social or material entities, while others, such as Pickering (1993), even discuss shifts in the agency of these entities. Some scholars, like Law (1992), go further, arguing that such entanglements can even impact the existence of the entities themselves (e.g., the creation of technological artifacts or new roles for humans). In essence, this notion illustrates the ways in which imbrication shapes both entities (Leonardi, 2012). *Strong relationality* implies that significant changes in the properties and agency of social and material entities arise directly from their interaction (M. Jones, 2014). For example, in recruitment processes, AI systems that use neutral, objective criteria to evaluate candidates can significantly alter how human recruiters approach their own decision-making (Zheng et al., 2024). As the AI evaluates applicants based on ethical standards free from bias, human recruiters may gradually adopt more objective and impartial behaviors themselves, reshaping their own ethical norms and practices (Fritts & Cabrera, 2021; Zheng et al., 2024). In this case, AI's neutral criteria enhance the ethical behavior of the human recruiter, resulting in a co-evolution where both entities—AI and humans—impact each other's ethical perspectives and behaviors. The social actor (recruiter) is not simply aided by the AI but transforms their own evaluative process in response to the AI's objective analysis (Zheng et al., 2024). In comparison, *weak relationality* suggests that the interaction between social and material entities has minimal impact on their inherent forms, attributes, and capabilities, which remain largely independent of the relationship (M. Jones, 2014). For example, in the context of medical AI diagnoses, moral judgments about responsibility may not significantly change (Huo et al., 2022). Patients' self-responsibility attribution enhances acceptance of medical AI, even though misdiagnoses can occur (Huo et al., 2022). Or, in a hiring context, recruiters still perceive AI agents as neutral, although there have been moral violations within the system (Hong et al., 2020). In those cases, the ethical and moral values of the human participants remain intact, with AI treated similarly to human counterparts in ethical judgments. This minimal change in moral agency or ethical behavior illustrates how, despite the presence of AI, the core attributes of social entities—such as their ethical framework or virtue ethics—remain unaffected. Thus, the sociomaterial notion of relationality can be used to analyze how the interaction between humans and technologies can lead to changes in either the AI's or the human user's properties (such as ethical capabilities), agency (such as moral values or ethical principles), or existence (such as ethical roles).

Performativity is a sociomaterial notion that underscores the interdependencies between an organization's social subsystem and humans as social entities engaged in sociomaterial practices (Ceccez-Kecmanovic et al., 2014). It elucidates how individuals construct their reality through social interactions with material entities, shaped by the enactment of boundaries within organizational relationships (M.

Jones, 2014). Orlikowski and Scott (2008, p. 462) assert that “the notion of performativity draws attention to how relations and boundaries between humans and technologies are not pre-given or fixed, but enacted in practice.” *Strong performativity* suggests that the boundaries and relationships—such as technological constraints, human mental models, or social subsystem restrictions—shape perceptions of reality in sociomaterial practices, reflecting the natural world (M. Jones, 2014). For instance, strong performativity linked to ethics can be observed when AI systems impose limitations on decision-making, compelling humans to confront ethical dilemmas shaped by the technology’s constraints. This can be seen in the case of autonomous vehicles, where human decision-making authority and manual skills may deteriorate, creating a disconnect between responsibility and control over the system (Holford, 2022). In contrast, *weak performativity* posits that reality is shaped regardless of the boundaries and relations of nonhuman agencies (M. Jones, 2014). An example of weak performativity linked to ethics could involve AI’s role in ethical decisions, where human users are not bound by predefined technological limits but are influenced by AI suggestions. For instance, in complex traffic scenarios with unavoidable accidents, even though the autonomous car can make recommendations, human drivers may have to make decisions largely instinctual, leading to moral shortcuts (Cunneen et al., 2019). Thus, the sociomaterial notion of performativity can be used to analyze how boundaries, limitations, and beliefs regarding what is considered “right” or “wrong” in the workplace as a social subsystem (incorporating various normative ethical perspectives), the human user’s values (virtue ethics), and the AI’s parameters (deontology) collectively shape how individuals construct their reality and moral perceptions.

Another sociomaterial notion is *materiality*. Materiality generally refers to an old and well-established research stream within IS and organizational science literature, which examines the material aspects of technological artifacts and their properties—those that “do not change from one moment to the next or across differences in locations” (Leonardi, 2012, p. 32). Jones (2014, p. 987) critiqued the inconsistent usage of the term, stating that “materiality is variously identified with the material, artefacts, the tangible, machine, nonhuman, and technology. ... Treating all these terms as equivalent, therefore, would seem to be a source of potential inconsistency in discussions of sociomateriality.” In organizations, materiality pertains to how technologies are arranged as relatively stable and effective components of social practices, whereby these material components can include both tangible objects, such as hardware, and intangible ones, such as data and algorithms, which are embedded in physical artifacts (M. Jones, 2014; Leonardi, 2012). Thus, materiality illustrates a connection between the social subsystem and the material entity (Leonardi, 2012). In the context of human-AI interaction, materiality focuses on how AI, as a material entity, is organized and arranged within specific artifacts (e.g., within particular smartphone applications) and its relationship with other material objects such as ethical guidelines for AI usage, which are conveyed as tangible documents within the social subsystem (Heyder et al., 2023). *Strong materiality* refers to the materialization of a phenomenon, indicating a robust integration of material entities within both the physical and non-physical components of an artifact, such as AI in robot-assisted surgery (M. Jones, 2014). For instance, strong materiality is evident in the context

of using AI-driven digital twins for smart manufacturing (Huang et al., 2021). These systems, embedded in physical infrastructure like cameras and sensors, are governed by fixed algorithms that trigger alerts based on predefined patterns (deontology). Ethical principles are materialized in these systems, driving rigid decision-making frameworks to the shape manufacturing processes and human-robot collaborations to ensure safety and data privacy (Othman & Yang, 2023). In contrast, *weak materiality* refers to a more distinct arrangement of the material entity (e.g., generative AI applications) and other artifacts that are related to the material entity but remain relatively stable and unchanging across different locations and time periods (M. Jones, 2014). An example of weak materiality is guidelines for AI design and utilization to better address ethical issues (Wang et al., 2020). Even though a company may change a specific AI system for another (thus, the material entity changes), the guidelines for AI design and usage based on deontological principles that the organization claims may remain. Thus, the sociomaterial concept of materiality can be used to analyze how AI technology's ethical frameworks—such as its deontological principles (programmed rules and constraints), virtue ethics (the moral values of human stakeholders, including employees and decision-makers), and consequentialism (the outcomes and consequences for the organization)—are shaped by the arrangement and integration of AI within artifacts. These interrelations determine how AI systems influence the construction of ethical norms and behaviors within the organization.

Finally, *practices* illustrate another sociomaterial notion that involves the enactment of performativity, which refers to how routinized and affective activities and processes are enacted in real-world contexts (Heyder et al., 2023; M. Jones, 2014). This highlights the way in which the entanglement of social and material entities influences both affective practices arising from workforce emotions and routine practices based on structures and guidelines from the social subsystem (Heyder et al., 2023; M. Jones, 2014). *Strong practices* demonstrate that the material becomes an integral part of these activities by shaping specific routinized and emotional responses, while *weak practices* suggest that activities and routines remain unchanged, as the entanglement does not significantly alter how these processes are conducted in the real world (M. Jones, 2014). An example of strong practices is provided by a study of Uber drivers, where researchers found that drivers subjected to algorithmically managed platforms often experience frustration as a strong emotional response to perceived unfairness and lack of control when receiving low ratings from passengers, knowing that such ratings could possibly lead to account deactivation (M. K. Lee, 2018). To avoid such scenarios, drivers tend to behave opportunistically, attempting to overcome the platform's ranking by identifying loopholes (M. K. Lee, 2018). This illustrates how perceived unfairness and the lack of control enforced through the platform's strict parameters (linked to deontological principles) might clash with the driver's virtuous character and perceptions (linked to virtue ethics), leading to some forms of affective practices that are opportunistic and, thus, nonethical activities. In comparison, an example of weak practices would be an organization implementing a decision-support system for managerial guidance in performance evaluations without altering the existing human-driven processes. Although the AI might provide recommendations, the

introduction does not fundamentally change daily practices (Hammerschmidt, Stolz et al., 2025). Thus, the sociomaterial notion of practices can be used to analyze how AI transformation alters human interactions with AI technologies (sociomaterial entanglement), influencing established routines (e.g., ethical guidelines grounded in deontology) and certain working practices that reflect the emotional and ethical dimensions of the workforce within the social subsystem and vice versa.

In summary, the weak and strong notions of sociomateriality, which represent distinct ontological positions, can be used as a valuable lens for explaining and understanding perceptions of ethics and moral agency within human-AI interactions. This extends the conceptualization of Heyder et al. (2023) by integrating multiple normative ethical theories and focusing on the shifting dynamics of moral agencies between AI and humans. The sociomaterial notions enhance our understanding of ethics within human-AI interaction, emphasizing their close connection to sociomaterial entanglement, as illustrated in the results of this dissertation (see Section 4.1). Meanwhile, these notions further offer insights into how managerial strategies and processes within the social subsystem influence ethics in the context of sociomaterial entanglement and vice versa, contributing to the results in Section 4.2. Together, they form a sociomaterial cycle, meaning that all notions of sociomateriality are interconnected; for instance, inseparability is closely connected to relationality, and performativity influences inseparability (Heyder et al., 2023). Thus, utilizing Jones' (2014) weak and strong sociomaterial notions and integrating them into Leonardi's (2012) conceptualizations of sociotechnical systems, as demonstrated by Heyder et al. (2023), provides a valuable lens for investigating the understanding and management of ethics within human-AI interactions. However, it is important to note that sociomaterial scholars do not view those notions as definitive or exhaustive. As Jones (2014, p. 921) cautions, these notions are

...not intended to suggest that they constitute a complete account of all possible ways in which sociomateriality is, or might be, constructed, or that they represent the opposite ends of a single spectrum on which all other variants might be located.

Thus, this dissertation aims to address a gap in sociomateriality literature by exploring the different ways sociomaterial notions can be combined in an ethical context, as requested by Jones (2014).

Figure 1 provides a conceptualization of the interplay between ethical perspectives of the entities within an organization as sociotechnical systems. This interplay is illustrated through the sociomaterial notions, highlighting that the allocation of the different ethical perspectives is not complete and exclusive. The conceptualization extends the work of Heyder et al. (2023).

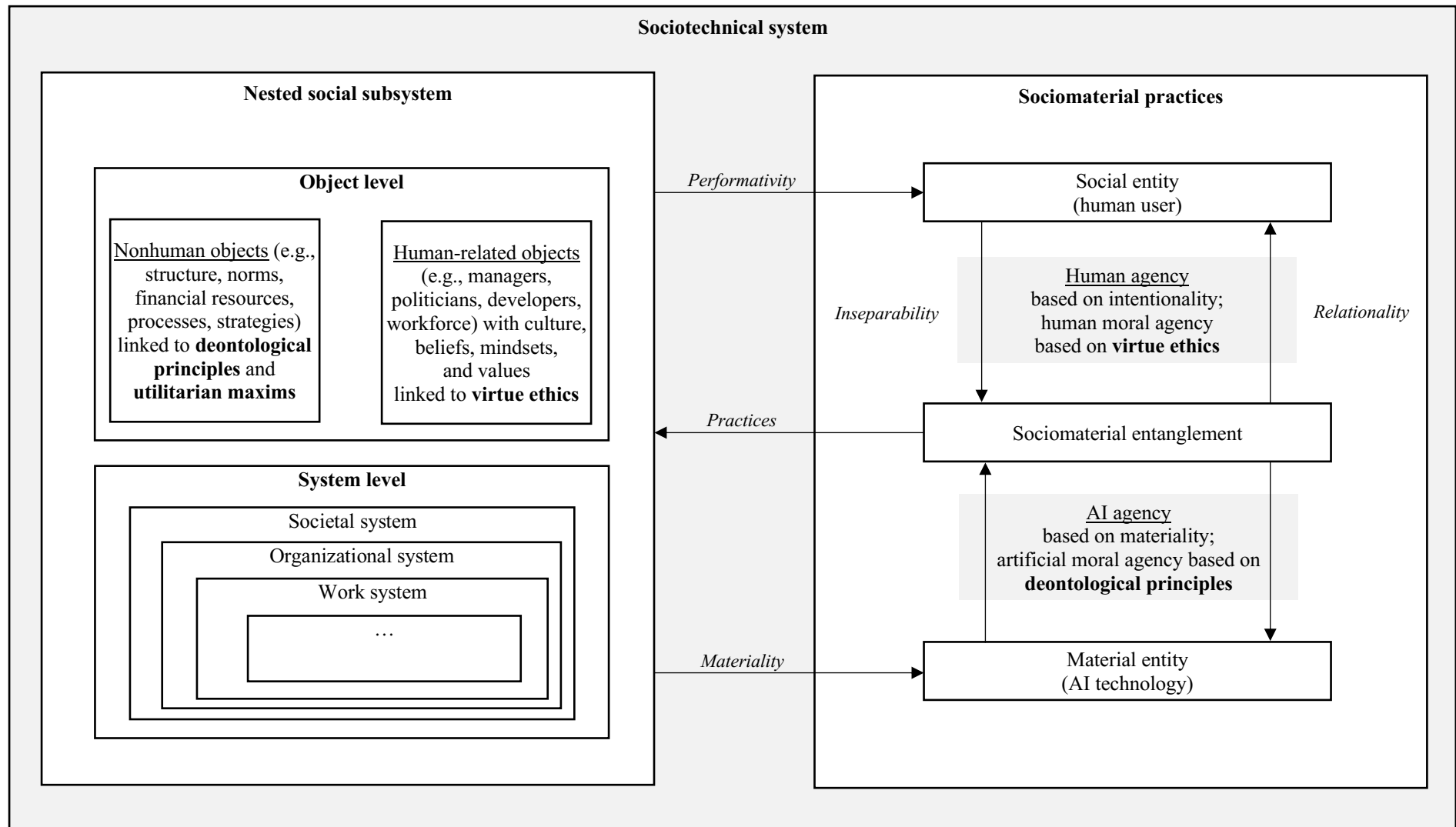


Figure 1: Conceptualization of the interplay of ethical perspectives in a sociotechnical system illustrated through the sociomaterial notions (extends the work of Heyder et al. (2023)).

3 Methodology

As outlined in the introduction, this dissertation combines systematic reviews to synthesize, describe, and explain previous insights into how we can understand ethics and ethical management in human-AI interactions, based on well-established research (mainly deductive), with insights from primarily exploratory empirical investigations to gain deeper insights into specific patterns identified in the literature reviews (mainly inductive). In doing so, this dissertation integrates deductive insights from systematic literature reviews, building upon well-established research (Paré et al., 2015; Templier & Paré, 2018), with inductive insights from empirical investigations (Sarker et al., 2018). The sociomaterial notions integrated into sociotechnical system theory are employed to analyze the changing nature of human-AI interaction, particularly AI's increasing agency and its impact on perceptions of moral agency in both humans and AI, a new phenomenon in the IS field. By combining well-established ethical theories with empirical research on the abilities and requirements for individuals, organizations, and technological artifacts to interact ethically with AI technologies, this dissertation seeks to enhance our understanding of how these perceptions and abilities shape ethical human-AI interaction. Additionally, by illustrating how sociomaterial notions can be utilized to represent different ethical perspectives within organizational systems, this dissertation aims to increase knowledge of how to manage human-AI interaction ethically. This dissertation extends established theories by showing how ethical perspectives can be applied within organizations as sociotechnical systems to provide a structured approach to investigating the changing dynamics between human and AI agency in organizations. This aligns with the *extend strategy* of “Next-Generation Theorizing” of Burton-Jones et al. (2021, p. 303) by combining established ethical theories and sociomaterial conceptualizations to study emerging phenomena in the IS field.

The nature of theory in this dissertation is primarily explanatory, focusing on “how, when, where, and why events occurred may be presented, giving rise to process-type theory” (Gregor, 2006, p. 624). The central research question explores how ethical human-AI interactions can be understood and managed. Sociomaterial notions are used as a theory for analysis, while ethical theories provide explanatory frameworks (Gregor, 2006). Both offer research questions that explore how ethical human-AI interactions can be understood and managed. Through empirical insights, this dissertation aims to explain “how and why things happen in some particular real-world situations” (Gregor, 2006, p. 624). By combining interpretivist approaches (via empirical insights) with logical positivist approaches (through established theories that integrate substantial and relational ontologies), this work contributes to a deeper understanding of ethical human-AI interaction.

This dissertation presents a research approach that aligns with the four key shifts in theorizing outlined by Burton-Jones et al. (2021, p. 306). *First*, by building on the discourse of sociomateriality, this dissertation aims to “appreciate the history of ideas in IS” while integrating new perspectives on the

changing dynamics between social and material (moral) agencies in human-AI interaction (Burton-Jones et al., 2021, p. 306). This approach reflects an understanding of theory as an “intellectual conversation” between established and emerging scholars, both within IS and across fields such as moral philosophy (Burton-Jones et al., 2021, p. 306). In this way, established contributions are acknowledged, while certain assumptions—such as the view of material agency as passive and reactive only to human input (e.g., Leonardi, 2011, 2012)—are challenged, given AI’s increasing capabilities for unsupervised learning and semiautonomous decision-making. *Second*, this dissertation views “[t]heory as a lens,” recognizing the diverse ontological choices in previous research (Burton-Jones et al., 2021, p. 306). It emphasizes the importance of reflecting on how a study’s methods and conceptual frameworks represent a specific lens. The complexity of AI’s interaction with humans, given the wide range of AI use cases and their impact on ethical perceptions (Figueras et al., 2022), necessitates a nuanced understanding. Thus, sociomateriality is used as a lens, alongside normative ethical theories, to analyze the ethical dimensions of human-AI interaction. Both are connected to different ontological positions, highlighting the complexity of how we construct ethical frameworks and understand human-AI interactions. *Third*, by combining systematic literature reviews with empirical investigations, this dissertation embraces theory as a “product of reflective engagement and action research” (Burton-Jones et al., 2021, p. 306). The integration of established conceptual knowledge with new field insights allows for the generation of “causal explanations” (Gregor, 2006, p. 630), blending desk-based theorizing with practical insights from real-world practice. *Fourth*, this dissertation acknowledges that the sociomaterial phenomena being studied are themselves shaped by the theorizing process (Burton-Jones et al., 2021). In examining how to ethically manage human-AI interactions, the research remains mindful of the performative nature of its theoretical lens, recognizing that the way the phenomenon is framed can influence real-world outcomes, both positively and negatively.

In the subsequent section, this dissertation offers an overview of the main procedures used in the systematic literature reviews and empirical investigations conducted in its articles. The methodological approaches are primarily based on Templier and Paré’s (2018) guidelines for increasing the transparency of literature reviews and on Sarker et al.’s (2018) recommendations for enhancing the rigor of qualitative research.

3.1 Systematic Literature Reviews (Primarily Deductive)

This dissertation comprises five articles that employ systematic and stand-alone literature reviews (**Papers I, II, VI, VII, and VIII**), with the aim of describing and explaining patterns, interdependencies among constructs, and trends within the existing body of literature. These reviews use a structured and systematic approach to reorganize and synthesize constructs from prior research, thereby generating novel insights (Boell & Cecez-Kecmanovic, 2015; Kitchenham et al., 2009). The dissertation employs various types of literature reviews for different purposes. In line with Paré (2015), **Paper II** utilizes a

scoping review to map key concepts within a research field, assessing its nature, trends, extent, scope, and gaps. **Papers I and II** adopt a descriptive review approach, aiming to provide a general understanding of a topic by identifying and describing patterns in existing literature (Paré et al., 2015). Theoretical reviews are employed in **Papers VI and VII** to expand previous conceptualizations within a specific field and to propose new theoretical perspectives (Paré et al., 2015).

All studies followed the systematic review process outlined by Templier and Paré (2018), which involves three main steps:

The first step involved developing and conducting a *search strategy*, which included constructing keyword-based search strings across multiple databases to ensure broad coverage of interdisciplinary literature. This included initial queries and pilot searches to identify relevant databases and refine the search string for each article (Kitchenham et al., 2009). Commonly used databases included the ACM Library (ACM), AIS Electronic Library (AIS), Business Source Ultimatum via EBSCOhost (EBSCO), and Web of Science (WoS), as these databases cover major academic journals and conferences from the IS domain, as well as those in the fields of ethics, moral philosophy, management, and organizational science (Heyder et al., 2023). To capture a broader range of relevant literature, the search strings often included abstract terms such as “AI” and “ethics,” which increased the volume of results but necessitated additional effort during the screening process (Boell & Cecez-Kecmanovic, 2015). Once the final search was completed, duplicates were removed to identify the unique hits that remained for the next step. For **Papers VI and VIII**, backward and forward citation searches were additionally conducted to identify further relevant articles not covered by the databases.

The second step involved the *literature screening*, which included formal, relevant, and—when necessary—quality screenings based on clearly defined inclusion and exclusion criteria. All reviews began with a format screening, excluding non-English literature, stand-alone literature reviews without new contributions, research-in-progress articles, and gray literature. Database filter options were employed wherever possible to streamline this initial phase. While quality criteria were generally linked to peer-review standards, this dissertation often employed less stringent quality thresholds to balance the inclusion of interdisciplinary research streams and encourage theoretical diversity (Markus & Rowe, 2021). However, **Paper VI** involved a more rigorous quality screening, using journal and conference rankings to restrict the review to high-quality publications, applying established ratings such as the VHB JOURQUAL Rating (VHB, 2015). All reviews underwent multiple rounds of relevance screening, involving substantial discussions among coders to enhance coding reliability. For most papers, interrater reliability was measured using Cohen’s Kappa following Cohen (1960) to ensure consistency (except for **Paper VII**, which was a single-author publication) (McHuhg, 2012). This procedure resulted in a final selection of literature, which was downloaded as PDFs.

In the final step, a *qualitative analysis* was performed on the final sample. Each article was analyzed using deductive coding based on the coding scheme of Mayring (2014, p. 95), which included coding definitions, examples, and rules for each coding category. The qualitative analysis was primarily conducted using MAXQDA software, and the coding scheme was refined through iterative discussions and coder training sessions (Mayring, 2014, pp. 56–62). The deductive coding process followed the steps outlined by Mayring (2014, p. 95), starting with defining the research question and category system, followed by establishing coding guidelines, conducting a material run-through, revising categories, and finally coding the entire sample, leading to an analysis of frequencies and interpretations. Meta-information was also extracted from the final sample to provide a descriptive overview of key characteristics, such as methodology and theoretical frameworks.

To ensure the quality of the review, all the review articles of this dissertation have a systematic nature “as a methodology for performing unbiased aggregation of empirical results” (Kitchenham et al., 2011, p. 638). The steps of Templier and Paré (2018) were followed with the aim of enhancing transparency. The coding process was strengthened through iterative rounds of coding and extensive coder discussions, following Mayring’s (2014, pp. 56–62) guidelines to ensure rigor in the qualitative analysis.

3.2 Empirical Investigations (Primarily Inductive)

Five of the articles in this dissertation are based on empirical investigations (**Papers III, IV, V, IX, and X**) aimed at exploring specific phenomena and patterns identified in the literature reviews that require further examination. A qualitative approach was chosen to investigate diverse ethical perspectives and moral values held by individuals, seeking to capture their subjective understanding of ethics in the context of human-AI interaction. The ability of qualitative approaches to capture subjective understandings of participants allows us to explore how ethical perceptions are shaped through social processes involving both the technology and the organizational environment (Sarker et al., 2018). These exploratory case studies focus on complex interventions, such as ethical interventions, across multiple settings (Yin, 1994). Most of the dissertation’s empirical articles utilize qualitative methods, specifically semi-structured interviews (**Papers V and IX**) and qualitative online questionnaires (**Papers III and X**), whereby **Paper III** additionally uses two closed-item scales to classify participants’ qualitative answers into different groups regarding their levels of AI literacy and emotional attitudes toward AI. The qualitative data provide insights into individual beliefs and opinions closely aligned with virtue ethics, particularly regarding human-AI interaction, AI agency, and ethical considerations. Thus, the short paper (**Paper V**) additionally applies a theoretical review before conducting interviews following a taxonomy development approach (Nickerson et al., 2013). **Paper IV**, in comparison, employs a quantitative approach through an online survey, followed by statistical analysis to test initial hypotheses derived from the literature. It also adopts an exploratory aim, examining how leaders’ capabilities may

influence organizational strategies and, consequently, how organizations develop the capacity to manage AI technologies effectively.

The empirical studies in this dissertation followed a consistent methodological process comprising three key steps:

The first step involved identifying appropriate *sample selection criteria* and a *sampling strategy*. To ensure a diverse cross-section of participants from various cultural backgrounds and relevant use cases of human-AI interaction, participants were selected from different stakeholder groups such as managers, AI developers, and educators. At least five participants were included in each group to ensure that collective insights emerged beyond individual perspectives. Each stakeholder group consisted of individuals with a minimum of three years of experience and expertise in their respective fields. Experts were primarily identified through social media platforms, particularly LinkedIn Premium profiles, and contacted via private messaging, which sometimes transitioned to email correspondence (particularly **Papers V and IX**). **Papers III and X** additionally used the crowdsourcing platform Clickworker, which enabled more targeted participant recruitment and helped mitigate biases associated with relying on a single platform (Aguinis et al., 2021). This selection approach ensured a high degree of participant heterogeneity (T. W. Lee et al., 1999). As these heterogeneous samples were recruited through diverse channels, they represented a convenient sampling method (Etikan, 2016). To enhance the quality of the sample (especially for online questionnaires), aptitude checks were included in the survey (e.g., questions about participants' prior experience with AI), and data from participants who did not pass these checks were excluded from further analysis. Across all papers, LinkedIn provided the highest quality responses, although it had the lowest response rates compared to crowdsourcing platforms (Hammerschmidt et al. 2024; Stolz et al., 2024).

The second step involves the *questionnaire design*. Most articles employed semi-structured interview guides with predefined sections and questions to ensure a standardized procedure (Myers & Newman, 2007). **Papers III and X** were based on qualitative online questionnaires featuring open-ended text fields, while **Papers V and IX** involved personal interviews conducted and recorded via Zoom. These interviews were later transcribed for analysis. While standardized questions were used for all participants, the semi-structured nature of the guide allowed for follow-up questions when participants' responses warranted deeper investigation (Myers & Newman, 2007). The exploratory focus of these qualitative interviews enabled a detailed investigation into ethics-related opinions and beliefs from each participant's perspective (Creswell, 2007, pp. 73–75). **Paper IV**, in contrast, relied on a quantitative online questionnaire with closed-ended items drawn from well-established scales in previous research.

The final step is about *data analysis*. For the qualitative studies (**Papers III, V, IX, and V**), data analysis was primarily conducted using MAXQDA software for qualitative content analysis following

Mayring's (2014) approach. Unlike the literature reviews, the qualitative data focused more on inductive coding. Multiple coding rounds were conducted, with intercoder reliability measured using Cohen's Kappa to ensure coding consistency (Cohen, 1960; McHuhg, 2012). The inductive coding process followed the steps outlined by Mayring (2014, p. 80), which include defining the research question and background, identifying potential categories and abstraction levels, refining categories and establishing coding rules through text analysis, further revising the categories and rules, final coding of the material, formulating main categories (if applicable), checking intercoder agreement, and interpreting the final results. For the quantitative analysis in **Paper IV**, statistical software was used to check for normal distribution, reliability, and multicollinearity, followed by multivariate regression analysis (Boudreau et al., 2001; Straub & Gefen, 2004).

To ensure the quality of the qualitative investigations (**Papers III, V, IX, and X**), each article followed "well-established 'first-generation' qualitative research approaches used in the discipline." (Sarker et al., 2018, p. 735). The quality of the interviews was enhanced by adhering to the qualitative interviewing guidelines provided by Myers and Newmann (2007). Similarly, the rigor of qualitative analysis was strengthened through the use of a coding scheme and iterative rounds of coding (Mayring, 2014, pp. 56–62). For the quantitative investigation in **Paper IV**, the article followed the validation guidelines for IS research provided by Straub and Gefen (2004) and adhered to the construct specifications outlined by Petter et al. (2007).

4 Main Research Results and Paper Contributions

As outlined in the introduction, the primary objective of this dissertation is to investigate the understanding and management of human-AI interaction from an ethical perspective, accounting for the changing dynamics between human and AI moral agency. By analyzing how different ontological positions shape our understanding of ethics and moral agency in human-AI interaction, resulting in ethical or unethical outcomes, this dissertation integrates well-established ethical theories into previous conceptualizations of sociomateriality. Section 4.1 addresses the first research question, "*How do human perceptions and the capabilities of both humans and AI influence ethical human-AI interaction?*" This section examines how human perceptions of AI and different human and AI capabilities lead to different interpretations of and requirements for ethics in the context of human-AI interaction. Answering this question creates the basis for addressing the second research question, "*How can we understand the ethical management of human-AI interaction?*" in Section 4.2, which examines how management can promote ethical outcomes and behaviors in human-AI interactions by integrating ethical considerations into managerial decision-making and strategies. Thus, this dissertation focuses on identifying how perceptions and capabilities shape ethical human-AI interaction and how ethically aligned management practices promote these interactions. Accordingly, this cumulative dissertation includes ten papers that draw on ethics-related research in the context of human-AI interaction. For each research question, there

are articles that provide theoretical foundations and articles that provide empirical findings. Although some of the papers contribute to both research questions, the first five articles in this dissertation (**Papers I to V**) primarily address the first research question (RQ1), while the second five articles (**Papers IV to X**) focus mainly on the second research question (RQ2). Table 3 includes the core results of this dissertation's articles in answering the research questions and provides an overview of how the sociomaterial notions in each article contribute to the goals of this dissertation.

RQ1	How do human perceptions and the capabilities of both humans and AI influence ethical human-AI interaction?	
Paper I	Theoretical foundations	• Paper I provides theoretical foundations for RQ1 by explaining that perceptions of AI and human moral agency—whether AI is seen as a moral agent or patient—are shaped by different ethical contexts (linked to ethical theories) and ontological views. Different perceptions of AI and human moral agency can create uncertainties about what and who determine the right and wrong behavior in human-AI interactions and thus influence the nature of ethical or unethical human-AI interactions. • It highlights the challenge of creating cohesive ethical frameworks for AI, questioning whether they should be tailored to specific AI capabilities (strong performativity) or shaped by public perceptions (weak performativity). • The notion of performativity helps to explain how human perceptions of AI's capabilities and moral agency shape ethical realities within human-AI interactions.
Paper II		• Paper II provides the theoretical foundations for RQ1 by explaining why combinations of ethical theories, particularly deontological and virtue ethics, are required to shape responsibility and AI governance practices for ethical human-AI interactions. • It highlights the challenge of developing AI governance and responsibility practices that balance the societal values of specific moral communities (strong practices) and ensure global inclusivity by being independent of cultural context (weak practices). • The notion of practices helps to explain how human perceptions and AI capabilities influence ethical human-AI interactions by shaping responsibility routines and AI governance practices.
Paper III	Empirical evidence	• Paper III contributes to RQ1 by providing empirical insights into how AI literacy and ambivalent emotional attitudes enhance human agency, particularly in their virtue ethics, by creating more realistic perceptions of AI, which contribute to ethical human-AI interactions. • It highlights the critical role of AI literacy in fostering humans' realistic perceptions of AI's capabilities (strong relationality) and stabilizing neutral emotional attitudes of humans (weak relationality), as both strengthen human agency. • The notion of relationality helps to explain how human perceptions and AI capabilities interact to shape ethical outcomes by focusing on the dynamic relationships between AI and human agency.
Paper IV		• Paper IV contributes to RQ1 by providing empirical insights into how leaders' AI literacy and ambidextrous decision-making directly shape the development of organizational capabilities that are crucial for fostering ethical human-AI interactions. • It highlights that leaders with higher AI knowledge can better balance the risks and benefits of AI, ensuring that well-developed capabilities align with AI's ethical deployment (strong materiality), while avoiding the underdevelopment of organizational AI capabilities (weak materiality). • The notion of materiality helps to explain how leaders' AI knowledge and decision-making capabilities influence the ethical deployment of AI within organizations.

Paper V		• Paper IV contributes to RQ1 by providing empirical insights into how higher levels of critical and sociocultural AI literacy allow individuals to act more independently of AI, strengthening their virtue ethics and enabling them to critically assess AI's outputs. AI-literate individuals are better equipped to recognize potential biases and societal implications of AI, which fosters more ethical human-AI interactions. • It highlights that individuals with higher literacy act more independently and critically (weak inseparability), fostering ethical interactions, while low literacy leads to an over-reliance on AI (strong inseparability), which risks unethical outcomes. • The notion of inseparability helps to explain how the interdependence between human actions and AI capabilities varies based on individuals' AI literacy.
RQ2	How can we understand the ethical management of human-AI interaction?	
Paper VI	Theoretical foundations	• Paper VI provides theoretical foundations for RQ2 by explaining how both guidelines and rules from the AI artifact and social subsystem (deontology) and individual and workforce moral values (virtue ethics) must be integrated into the ethical management of human-AI interaction. • It highlights the need to align AI's deontological principles with human virtue ethics, fostering transparency, responsibility, and ethical behavior through integrated organizational strategies that guide both human and AI actions. • The interplay of all sociomaterial notions within the sociomaterial cycle helps to explain the ethical management of human-AI interaction by revealing how managing the interplay of deontology and virtue ethics among various entities within an organization—across inseparability, relationality, performativity, practices, and materiality—is essential for ensuring ethical human-AI interactions.
Paper VII		• Paper VII provides theoretical foundations for RQ2 by demonstrating how the ethical management of human-AI interaction requires managerial practices that balance diverse ethical views to overcome moral dilemmas through both the top-down establishment of ethical principles in regulatory frameworks and the bottom-up fostering of moral deliberation. • It highlights the need for practices that integrate human moral agency, ensuring AI systems align with both institutional rules and human-centered ethical values (strong practices). Not striving for alignment (weak practices) would foster existing moral dilemmas and hinder ethical human-AI interaction. • The notion of practices helps to explain the importance of aligning top-down ethical standards with bottom-up moral values, allowing for diverse ethical perspectives and fostering a routine of ethical reflection and decision-making in human-AI interactions.
Paper VIII		• Paper VIII provides theoretical foundations for RQ2 by demonstrating how ethical management of human-AI interaction requires reducing inequalities in accessing and using AI, as these impede the creation of sustainable outcomes with AI and hinder ethical human-AI interaction. • It highlights the interdependencies between guidelines, education, and communication from the social subsystem and the AI's design that must be effectively aligned (strong materiality) to create sustainable outcomes; otherwise, inequalities could potentially widen the AI divide (weak materiality). • The notion of materiality helps to explain how the arrangement and design of technological artifacts within social contexts can either promote or hinder the creation of sustainable outcomes related to ethical or nonethical human-AI interactions.
Paper IX	Empirical evidence	• Paper IX contributes to RQ2 by providing empirical insights into how organizations can manage ethical dilemmas in human-AI interactions. It demonstrates how organizations deal with decision dilemmas when deploying AI, particularly between exploring AI's potential and managing its risks. • It highlights the need for ethical management to maintain a balance between routinized, AI-driven processes and opportunities for human reflection (strong practices), ensuring that AI-driven efficiencies do not override human-centered ethical considerations (weak practices).

		• The notion of practices helps to explain how the ethical management of human-AI interactions requires both standardized processes and individualized human input when integrating AI for business optimization (exploitation) or innovation (exploration).
Paper X		• Paper X contributes to RQ2 by providing empirical insights into the ethical management of human-AI interaction, which requires a balance between supporting users' ethical inclinations and encouraging reflection for those who may not prioritize ethical behavior. This is achieved through personalized AI nudges and ethically designed AI systems that foster both trust and long-term behavior change. • It highlights that ethically complaint AI systems are crucial to ensure the ethical management of nudging mechanisms that can modify human agency—either by reinforcing users' existing ethical behaviors (weak relationality) or prompting deeper behavioral change (strong relationality), depending on how AI systems engage with users. • The notion of relationality helps to explain how AI nudging mechanisms within human-AI interactions can be used to impact human agency and ethical behaviors.

Table 3: Overview of core results of each paper linked to the research questions.

4.1 Understanding How Perceptions and Capabilities Influence Ethical Human-AI Interaction

Regarding the first research question—how the capabilities of humans and AI, as well as human perceptions of AI, impact ethical human-AI interaction—this dissertation includes two articles that establish theoretical foundations through comprehensive literature reviews (**Papers I and II**) and three articles that provide empirical insights (**Papers III, IV, and V**). In the subsequent sections, there is a distinct description of each paper's contribution to answering RQ1. Although all sociomaterial notions are useful for analyzing ethical human-AI interaction in each of this dissertation's papers, the specific sociomaterial notion most suitable for explaining the contributions to RQ1 within a single paper was selected. Building on the sociomaterial notions, the following analysis illustrates how the interplay of different capabilities of humans and AI, as well as several perceptions of AI from various stakeholders (related to distinct ethical perspectives), can shape the sociomaterial entanglement within a sociotechnical system, thus determining whether human-AI interactions are ethical or not.

Theoretical Foundations for RQ1: Paper I (Performativity)

Paper I: Hammerschmidt, T.; Hafner, A.; Stolz, K.; Passlack, N.; Posegga, O. & Gerholz, K-H. (2025). A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation. Information Systems Frontiers (ISF), <https://doi.org/10.1007/s10796-025-10596-0>.

The first paper of this dissertation focuses on how the increased integration of AI technologies into business and social contexts creates moral uncertainties, particularly due to changing perceptions of agency that result in blurred responsibilities between humans and AI in decision-making roles. The motivation for this paper arises from the need to explain how different ethical perspectives shape our understanding of morality and responsibility in human-AI interactions and how these perspectives may help address or exacerbate the moral uncertainties caused by AI's changing agency, which intensifies responsibility gaps. This paper draws on well-established moral philosophies, including Kant's (1991)

categorical imperative (linked to deontology), Mill's (1989) and Bentham's (1988) work on utilitarianism, and Rawls' (1971) theory of justice. In addition to these normative theories, it provides background on meta-ethics, applied ethics, and descriptive ethics (e.g., Gert, 2005; Mackie, 1990; Singer, 1968) and further describes the differences between AI's moral and ethical agency (e.g., Bankins & Formosa, 2023; Behdadi & Munthe, 2020). It employs a descriptive literature review, using a keyword-based search strategy across eight databases from different research domains to ensure comprehensive coverage of ethical discussions surrounding AI. Out of 5,399 initial results, the final sample of 151 articles was selected for further bibliometric analysis, which mapped the author and citation networks.

Paper I's three major contributions to RQ1 are framed by the notion of *weak* and *strong performativity* as they relate to how individuals construct their ethical realities in sociomaterial practices shaped by the enactment of boundaries within an organization's social subsystem.

First, the paper outlines a dominance of deontological ethics in AI research and practice, which often overlooks other ethical perspectives such as virtue ethics and consequentialism. This creates ethical blind spots and leads to fragmented views on what constitutes ethical human-AI interaction, further illustrating a *strong performativity* since ethical dialogue is bound to regulation. Therefore, AI is perceived solely as rule-following agents, programmed to adhere to predefined ethical guidelines, and ethical behavior is defined by norm compliance. This limited view may encourage rigid adherence to legal frameworks, where sticking to regulations is seen as sufficient for ensuring ethical AI interactions. For instance, AI's strict adherence to rules could result in unethical outcomes if the rules themselves are flawed or biased, and the individual's reality (linked to their virtue ethics) is restricted to these rules, as the AI lacks the ability to assess the broader moral implications beyond those rules. In *weak performativity*, deontological ethics that fail to address the diverse moral frameworks that exist globally can create ethical misalignments in some communities that may hold higher ethical standards than those outlined in laws or regulations. For instance, in a collectivist community, ethical principles within an AI system might be designed to prioritize group welfare (consequentialism), while in an individualistic community, the focus lies on personal autonomy (virtue ethics). Thus, Western deontological frameworks might compete with Eastern ones, which illustrate *weak performativity*, where various cultural values are considered but without a deep integration of any one cultural or societal framework within AI regulation. Hence, when ethical dialogue and practices are bound to dominant ethical perspectives, such as deontology, this can—even though it is based on ethical principles—potentially hinder ethical human-AI interactions, which require balanced approaches of *weak* and *strong performativity*.

Second, the study underscores the challenges in creating cohesive ethical frameworks for AI that effectively contribute to ethical human-AI interaction. A key question raised by the paper is whether

these frameworks should be tailored to the specific functions and actual capabilities of different AI technologies (illustrating *strong performativity*), or whether AI should be treated as a unified entity shaped by public perceptions of AI capabilities (illustrating *weak performativity*). The findings suggest that while it is essential to develop tailored ethical frameworks that account for the unique capabilities and limitations of each AI system, it is equally important to ensure a unified ethical baseline. This is necessary to prevent tailored frameworks from falling below a minimum ethical and legal standard (linked to deontology) and to ensure that they align with public expectations (linked to virtue ethics). When ethical frameworks are closely tied to AI's capabilities and limitations, this represents *strong performativity*, as the ethical reality created for human-AI interaction is tightly linked to the boundaries of the technological artifact and its actual capabilities. Conversely, when ethical frameworks are not specialized to the unique capabilities of AI, it illustrates *weak performativity*, where ethical frameworks are shaped more by public perceptions of AI's potential and concerns than by the technical realities of the system. A framework that is too specialized may fail to meet broader ethical expectations, while a unified framework may not adequately address AI's limitations. Thus, developing ethical frameworks for human-AI interaction requires striking a balance between *strong performativity*—grounded in the actual capabilities and limitations of different AI technologies—and *weak performativity*—where public perceptions and concerns are addressed.

Third, the paper highlights different opinions among researchers regarding whether AI systems can exhibit moral character and how this affects our understanding of responsibility. These debates are often linked to varying ontological positions, which shape perceptions of both human and AI moral agency. The interaction between these perceptions and the environment can be explained through the sociomaterial notion of performativity, illustrating how different contexts shape the extent to which moral agency is attributed to AI. When AI systems demonstrate adherence to norms and surpass human morality in specific tasks, individuals may perceive these systems as possessing moral agency, attributing responsibility directly to the AI. This perception aligns with *weak performativity* and a substantial ontology, where the material capacities of AI are viewed as fixed and autonomous, existing alongside the social world but not fully integrated into it. In contrast, when AI's moral agency is seen as co-constructed through its interactions with human actors—learning from human data inputs, adapting to social contexts, and reflecting human intentions—this represents strong performativity grounded in a relational ontology. Here, AI's actions and ethical outcomes are viewed as relationally intertwined with the values, virtues, and intentions of the humans who design, deploy, and interact with the technology. AI, in this view, is not an independent moral agent but rather a moral patient, performing actions based on human moral frameworks. However, if AI is influenced by unethical users or designers, its actions may reflect these negative traits, resulting in unethical behavior. Therefore, the paper emphasizes the relevance of shared responsibility among various human stakeholders. While humans bear the ultimate responsibility for the outcomes of AI systems (representing *strong performativity*), this view risks overlooking the autonomous actions of AI and the ethical environments these systems can

create. Failing to recognize AI's capacity to act autonomously in certain contexts could lead to misplaced responsibility or an oversimplified understanding of AI's role in ethical human-AI interactions. Thus, finding a balance between *weak* and *strong performativity* is essential. Recognizing that AI's ethical behavior is co-created through its interactions with humans and the surrounding environment (*strong performativity*) is crucial, but it is equally important to acknowledge that AI can be misused by humans, necessitating a form of distinct moral agency tied to certain norms (*weak performativity*).

In conclusion, **Paper I** underscores that varying and bounded perceptions of ethical AI practices and frameworks, as well as different ontological positions on AI's moral agency, play a crucial role in determining ethical or unethical human-AI interactions. The paper establishes theoretical foundations for RQ1 by highlighting the importance of balancing normative ethical perspectives of different moral communities and further considering ethical frameworks that balance AI's actual capabilities and limitations with those perceived by the public. It further outlines the relevance of discussing both ontological positions of AI's moral agency (moral agent and moral patient) as this influences the nature of ethical or nonethical human-AI interactions in either weak or strong ways (referring to the sociomaterial notion of performativity).

Theoretical Foundations for RQ1: Paper II (Practices)

Paper II: Hafner, A.; Hammerschmidt, T. & Stolz, K. (2024). Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review. Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany, <https://aisel.aisnet.org/wi2024/46>.

The second paper of this dissertation aims to contribute to the field of corporate digital responsibility (CDR) by assessing how various normative ethical theories influence AI ethics practices. The motivation stems from addressing gaps in current AI ethics research, particularly in understanding the diverse ways ethical theories impact AI ethics practices and, thereby, ethical human-AI interactions. This specifically contributes to CDR, focusing on how organizational values can be aligned with certain norms to achieve responsible actions (Lobschat et al., 2021). The manuscript provides an overview of the most commonly applied normative ethical theories in AI ethics research: deontology, consequentialism, and virtue ethics (e.g., Bankins & Formosa, 2023; Brendel et al., 2021) and highlights the importance of understanding how different ethical theories shape perceptions of responsibility between humans and AI (e.g., Behdadi & Munthe, 2020; Coeckelbergh, 2020). A scoping review, drawing data from four IS-related databases, with 1,234 total hits and 61 articles in the final sample, was conducted to outline trends, research gaps, and the most studied topics regarding how ethical theories shape AI ethics practices.

Paper II's two major contributions to RQ1 are framed by the notion of *weak* and *strong practices* as it describes how AI governance routines and responsibility practices within the social subsystem are shaped by human-AI interactions.

First, the paper emphasizes the need for a more fundamental discussion on how our ethical understanding shapes responsible practices in human-AI interactions. It outlines that, currently, the dominance of deontology, which guides the design of AI systems to follow predefined ethical principles, may overlook the human user's moral responsibility, rooted in their virtue ethics, in ensuring ethical outcomes when working with AI. Deontological approaches tend to establish fixed routines for attributing responsibility based on normative rules, creating *strong practices* linked to the AI artifact that do not fully consider users' responsibility judgments based on their individual moral character. The underrepresentation of virtue ethics in human-AI interaction can lead to *weak practices*, indicating that current frameworks guiding AI ethics practices may inadequately address human roles in upholding ethical standards in human-AI interactions. Therefore, expanding the focus to include virtue ethics would balance *weak* and *strong practices* between the AI artifact and human users, resulting in a more balanced understanding of shared responsibilities within human-AI interactions.

Second, the paper argues that we need more empirical insights into how different value systems (linked to virtue ethics) impact AI governance practices, particularly in establishing ethical principles (related to deontology). Current regulatory and governance frameworks for AI, such as the European AI act, reflect the societal norms and values of specific cultures. When AI regulation is closely aligned with the values of the nations adhering to these rules, this represents *strong practices*, where human virtue ethics help shape the norms within the social subsystem. However, this approach can result in uneven global AI governance landscapes, with some regions highly regulated and others less sensitized to ethical concerns. In organizations, such disparities could create barriers to implementing AI ethics, potentially hindering ethical human-AI interactions on a global scale. In contrast, governance mechanisms for AI that evolve independently of specific societal values may lack the flexibility to address cultural differences in perceptions of responsibility and, thereby, could fail to encourage moral responsibility among developers, users, and organizations. This reflects *weak practices*, where AI artifacts shape AI governance principles within the social subsystem that do not align with the broader societal norms and values of that same subsystem. Such *weak practices* risk creating a one-size-fits-all AI governance framework that may not be sensitive to the ethical nuances of different regions. Therefore, balancing *weak* and *strong practices* can help establish AI governance frameworks that are globally inclusive while remaining flexible enough to account for cultural and value-based considerations, which can contribute to ethical human-AI interactions across diverse contexts.

In conclusion, **Paper II** contributes to understanding ethical human-AI interaction (RQ1) by outlining how AI ethics and responsibility practices within the social subsystem are shaped by human-AI interactions. Therefore, ethical theories should not only shape organizations' practices regarding how responsibility is attributed between humans and AI but also guide the development of governance practices that ensure AI systems are designed, deployed, and used in responsible ways. Thus, balancing

between *weak* and *strong practices* formed by the AI artifact and those formed by human users is crucial for ethical human-AI interaction.

Empirical Evidence for RQ1: Paper III (Relationality)

Paper III: Hammerschmidt, T.; Passlack, N.; Stolz, K. & Posegga, O. (2024). AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants? Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, <https://hdl.handle.net/10125/106426>.

The third paper of this dissertation advances our understanding of ethical human-AI interaction by exploring how individuals' perceptions of AI, particularly in relation to ethical AI narratives when using AI-based digital assistants such as ChatGPT or Alexa, are linked to their AI literacy and emotional attitudes toward AI in human-AI interactions. The study is motivated by the increasing role of AI assistants in shaping human behavior and decision-making during interactions, highlighting that incomplete or inaccurate understandings of AI can lead to unrealistic expectations and potential misuse (Cave & Dihal, 2019; Diederich et al., 2022; Fast & Horvitz, 2017). The paper explores how individuals' AI literacy—defined as the knowledge, skills, and competencies related to AI—and their emotional attitudes toward AI impact AI perceptions that contribute to specific behaviors in human-AI interactions. It builds upon the nine perception categories outlined by Diederich et al. (2022), extending them with insights from current literature on AI narratives (e.g., Cave & Dihal, 2019; Fast & Horvitz, 2017), ranging from optimistic (“utopian”) views of AI, such as AI enhancing productivity, to pessimistic (“dystopian”) concerns, such as AI replacing jobs. These polarized narratives, shaped by cultural, social, and media influences, affect how individuals perceive AI's role and how they interact with the technology. Additionally, the paper incorporates previous work on AI literacy (e.g., Long & Magerko, 2020; Pinski & Benlian, 2023) and emotional attitudes toward AI (e.g., Maier et al., 2019). Through an exploratory qualitative online questionnaire completed by 151 participants, the study clustered participants into different groups based on participants' emotional attitudes (positive, neutral, and negative) and varying levels of AI literacy (high, medium, and low). The analysis explores how these groups differ in their AI perceptions, as the experiences they have already gained with AI technologies create certain AI narratives, resulting in ethical or unethical behavior.

Paper III's two key contributions to RQ1 are framed by the notion of *weak* and *strong relationality*, as it emphasizes how human-AI interaction impacts human perceptions and ethical behavior.

First, the study offers valuable insight into the relationship between AI literacy and the development of a realistic perception that enhances human agency, thereby contributing to ethical human-AI interactions. The results indicate that when participants exhibit high levels of AI literacy, they are more likely to maintain realistic perceptions of AI agency. They understand that AI lacks human qualities, such as empathy and ethical reasoning, and are aware of the ethical implications associated with AI

usage (thus demonstrating increased virtue ethics). These individuals are more inclined to allow their human agency to guide the interaction, as they recognize AI's limitations and acknowledge the necessity of human oversight and robust data governance guidelines in AI-driven processes. This demonstrates *strong relationality*, wherein the ethical behavior and outcomes of the human-AI interaction are contingent upon human agency due to high literacy. It empowers them to critically assess AI's capabilities in decision-making processes while being aware of their own role in maintaining ethical standards by recognizing both the opportunities and risks associated with AI. In doing so, AI literacy strengthens their virtue ethics. In contrast, when participants demonstrate low AI literacy, they tend to hold unrealistic perceptions of AI agency, such as believing that AI can feel emotions or make moral judgments. Such scenarios reflect the *weak relationality* of the social entity, where these misunderstandings and the limited human agency exist irrespective of the interaction. Individuals may over-rely on AI, permitting it to make decisions that would necessitate human oversight. Therefore, AI agencies gain increased power due to human misconceptions, which can result in biased or harmful decisions when individuals fail to adequately monitor or intervene in AI-driven processes due to a lack of virtue ethics. Thus, the findings emphasize that AI literacy is crucial for fostering ethical human-AI interactions.

Second, the manuscript offers valuable insight into the relationship between emotional attitudes and the development of realistic perceptions that enhance human agency, thereby contributing to ethical human-AI interactions. The results indicate that participants exhibiting overly positive emotional attitudes (strong relationality), characterized by optimistic narratives—such as the belief that AI can relieve humans of mundane tasks or improve health outcomes—may develop an uncritical acceptance of AI's decisions. This can lead to a diminished capacity to engage critically with the technology (illustrating limited virtue ethics). Similarly, overly negative emotional attitudes can inhibit interaction with AI technologies in general due to a lack of trust in the technology (strong relationality). Participants with more neutral emotional attitudes (weak relationality) tend to have more cautious and critical perspectives toward AI by making more nuanced distinctions between artificial and human intelligence, enabling them to recognize AI's limitations based on strong virtue ethics. Therefore, the study argues that neutral emotional attitudes (weak relationality)—which incorporate both positive and negative perceptions—are essential for forming a realistic understanding of AI capabilities, as this enhances human agency in ways that support ethical human-AI interactions.

In conclusion, **Paper III** provides empirical insights into the understanding of how human capabilities (linked to AI literacy) and their emotional attitudes can impact the development of realistic AI perceptions, thereby strengthening human agency and virtue ethics based on sociomaterial entanglement. AI-literate humans with balanced emotional attitudes are better equipped to strengthen their human agency when interacting with AI technologies, thus contributing to ethical human-AI interaction (RQ1). Therefore, ensuring *strong relationality* by increasing humans' AI literacy and *weak*

relationality by stabilizing their neutral emotional attitudes through experiences with AI technologies is crucial for ethical human-AI interaction.

Empirical Evidence for RQ1: Paper IV (Materiality)

Paper IV: Hammerschmidt, T.; Stolz, K. & Posegga, O. (2024). How Leaders' Ambidexterity and Literacy Impact AI Capabilities of Organizations. Proceedings of 32nd the European Conference on Information Systems (ECIS), Paphos, Cyprus, https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14/.

The fourth paper of this dissertation addresses the need to understand how leaders' capabilities, particularly their ambidexterity—the ability to balance exploration and exploitation when making strategic decisions—affect an organization's capacity to develop appropriate capabilities for AI deployment and utilization. Given AI's increasing role in organizational strategies, leaders must make decisions that balance exploiting AI's efficiency and exploring its potential for innovation. This balancing act is critical for achieving business value from AI while managing its disruptive potential. The paper aims to fill a gap in understanding how leaders' knowledge and experience with AI (defined as AI literacy) influence their ambidextrous decision-making, which in turn affects both organizational ambidexterity and the development of AI capabilities. In doing so, this study offers valuable insights into how the interplay between AI capabilities and ambidextrous decision-making at both the individual (leadership) level and the organizational level may impact the ethical or unethical deployment of AI within organizations. The work builds on ambidexterity literature, particularly March's (1991) exploration-exploitation framework, employing different scales for measuring individual ambidexterity (Sirén et al., 2012) and organizational ambidexterity (Chen et al., 2022). It also draws on AI literacy measures from Pinski and Benlian (2023), distinguishing between AI knowledge (understanding of AI technologies, human-AI interaction, and AI processes), and AI experience (practical exposure to using and designing AI). Additionally, it incorporates Miklef and Gupta's (2021) conceptualization and operationalization of organizational AI capabilities, categorizing them into tangible resources (e.g., data management, technological infrastructure), human resources (e.g., technical and business skills), and intangible resources (e.g., interdepartmental coordination, change management, and risk proclivity). The study is based on a quantitative online survey of 216 leaders from German-speaking organizations.

Paper IV's two key contributions to RQ1 are framed by the notion of *weak* and *strong materiality*, as it explains the interdependencies on how leaders' ambidextrous decision-making capabilities and the development of organizational AI capabilities impact ethical human-AI interaction.

First, the paper's findings reveal that participants with higher levels of AI literacy—specifically, AI knowledge—are better equipped to make informed and balanced decisions about AI deployment within organizations. AI knowledge allows leaders to accurately anticipate the role of human actors in relation to AI technologies in human-AI collaborations, helping them mitigate potential risks. Understanding

how AI generates outputs and how humans should appropriately engage with these systems allows leaders to foresee potential pitfalls. These interdependencies can be illustrated through the sociomaterial notion of materiality, showing how ambidextrous leadership within the social subsystem influences how AI artifacts are arranged in organizations. Leaders lacking sufficient AI knowledge may either underutilize AI, thus missing opportunities to benefit from AI innovation, or over-rely on it, which illustrates limited virtue ethics that may lead to unintended negative consequences such as biased decision-making. Empirical evidence from the study demonstrates that lower levels of ambidextrous leadership are closely linked to corresponding deficiencies at the organizational level. Thus, in cases where participants are unable to strike a balance between exploration and exploitation, this imbalance hampers their organization's ability to cultivate the appropriate AI capabilities, which is a form of *weak materiality* as it cannot fully determine the extent to which human-AI interaction in organizations is ethical. In comparison, participants who have knowledge of AI's implications tend to carefully evaluate both the efficiency gains and the potential disruptive impacts of AI, thus demonstrating higher levels of individual ambidexterity, which correspond to higher levels of ambidexterity and capabilities at the organizational level. As a form of *strong materiality*, AI knowledge and capabilities of the social subsystem can contribute to more ethical AI deployment strategies by strengthening leaders' virtue ethics and, thereby, contributing to ethical human-AI interactions within organizations. Hence, the findings show that leaders with higher levels of AI knowledge are better equipped to make balanced decisions regarding AI deployment, resulting in ethical human-AI interactions by mitigating risks and leveraging AI's benefits similarly.

Second, the study highlights the importance of leaders' capabilities in balancing exploration and exploitation strategies for developing critical intangible and human resources that influence ethical AI deployment in organizations. Empirical findings demonstrate that exploration—focused on experimentation and innovation—significantly impacts the development of intangible resources such as change capacity and interdepartmental collaboration. Specifically, the higher a leader's emphasis on organizational exploration, the more likely the organization is to improve these intangible resources. The paper suggests that future research should consider additional intangible resources, such as organizational culture and workforce ethics, since these intangible factors closely linked to virtue ethics may play a more important role in shaping how organizations navigate exploration strategies. The findings suggest that developing robust intangible resources to effectively manage the risks associated with exploration strategies is crucial. Although intangible, these resources can create what can be understood as *strong materiality* within the social subsystem, meaning that well-developed intangible resources like collaboration and adaptability can shape how AI technologies are ethically deployed in organizations. Conversely, *weak materiality* arises when these resources are underdeveloped, which may lead organizations to pursue exploration strategies without adequately addressing the risks and unintended consequences associated with AI deployment. Empirical findings further indicate that exploitation—focused on optimizing existing operations and resources—has a significant impact on the

development of human resources, particularly employees' technical and business skills. The more leaders prioritize organizational exploitation, the more likely the organization is to enhance the AI-related competencies of its workforce. Since resources within organizations are often limited, leaders may focus on improving workflows and operational efficiency, recognizing the need to educate and sensitize employees to AI's impact and proper usage. When employees' skills are well developed, they know how to use AI technologies effectively within an organizational context, creating a *strong materiality* that supports ethical human-AI interactions. However, if these skills are underdeveloped, AI technologies may be embedded in the organization without sufficient oversight or control (*weak materiality*), which can hinder ethical AI deployment and lead to unintended negative consequences. Hence, the study underscores the critical role of intangible and human resources in shaping ethical AI deployment, demonstrating how leaders' focus on exploration and exploitation strategies significantly impacts the organization's ability to develop the necessary resources for ethical human-AI interactions.

In conclusion, **Paper IV** provides empirical insights into how leaders' AI literacy and ambidextrous decision-making capabilities influence the development of appropriate organizational AI capabilities for ethical human-AI interactions (RQ1). Illustrated by the sociomaterial notion of materiality, this paper considers how capabilities and strategies are formed through leaders' decisions within the social subsystem, and how this, in turn, might guide the sociomaterial entanglement, resulting in ethical or unethical human-AI interactions. Similar to **Papers III** and **V**, this work illustrates that the AI literacy of humans—whether they are leaders or employees—can strengthen one's virtue ethics and thereby contribute to more ethical interactions with AI.

Empirical Evidence for RQ1: Paper V (Inseparability)

Paper V: Heyder, T. & Posegga (2021). Extending the Foundations of AI Literacy (Short Paper, Best Paper Nominee). Proceedings of the 46nd International Conference on Information Systems (ICIS), Austin, Texas, https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/.

The fifth paper of this dissertation aims to extend the foundations of AI literacy in organizational settings. It argues that as AI technologies continue to advance and transform the workplace, employees must develop the skills to engage with AI systems responsibly. It aims to contribute to the creation of an AI-literate workforce, which increases our understanding of how AI literacy capabilities shape ethical human-AI interactions in several ways. It builds upon the definition of AI literacy by Long and Magerko (2020) and integrates existing work on digital, technological, and data literacy (e.g., Koltay, 2011; McLean, 2013) to establish a taxonomy of AI literacy with the following dimensions: functional AI literacy (e.g., knowledge of what AI is and how it works), critical AI literacy (e.g., evaluating and communicating the consequences of AI interactions and considering ethical implications), and sociocultural literacy (e.g., understanding certain cultural ideologies and the commonality of using AI). With the aim of developing a taxonomy of AI literacy, this study follows the methodology suggested by Nickerson et al. (2013) for taxonomy development, combining a systematic literature review with 42

overall hits and, in total, nine articles in the final sample in the conceptual phase, along with five semi-structured employee interviews in the empirical phase.

Paper V's two key contributions to RQ1 are framed by the notion of *weak* and *strong inseparability* as they illustrate interdependencies between humans and AI within the sociomaterial entanglement, highlighting more independent human actions when individuals have higher AI literacy capabilities.

First, the short paper extends previous conceptualizations of AI literacy by highlighting the importance of critical AI literacy in evaluating AI's impact and identifying potential biases in its outputs, which contributes to more ethical human-AI interactions. As one interviewee pointed out, when colleagues lack a solid understanding of AI due to limited technical knowledge, they often fail to grasp the challenges and ethical issues associated with AI. This leads them to either use the term "AI" as a buzzword, unaware that many of the technologies they use have underlying AI functionalities, or to make AI seem more tangible by describing it as human-like robots, attributing capabilities to the technology that it does not actually possess. Due to this lack of AI literacy, individuals either fail to recognize AI's presence during interactions or overestimate its capabilities, attributing more features to it than it realistically has. Hence, individuals with low levels of AI literacy are more dependent on AI's outcomes (reflecting *strong inseparability*), which restricts their virtue ethics in ways that can lead to unethical human-AI interactions, as they trust the technology uncritically. In contrast, for AI-literate individuals, this dynamic does not apply, as their capabilities strengthen their virtue ethics and make them more independent from AI's outputs (reflecting *weak inseparability*). These individuals can recognize AI's presence and critically assess its impact and potential biases, resulting in more ethical human-AI interactions.

Second, the short paper highlights the relevance of sociocultural AI literacy in addressing the broader societal implications of AI, which are crucial for ensuring ethical human-AI interactions. The preliminary findings suggest that data security concerns vary between private and work-related uses of AI technologies. In professional settings, individuals tend to exercise more caution when using AI, which illustrates *weak inseparability*, as humans (linked to their virtue ethics) deal carefully with AI technologies. In contrast, in their private lives, participants reported a more careless approach to using AI, which makes them more dependent on its outcomes (*strong inseparability*). Additionally, interviewees working in companies with a more homogenous workforce and a less regulated culture tended to be less critical of AI's output and less aware of AI's broader implications, reflecting *strong inseparability*. In contrast, those employed in organizations with a strong security orientation and diverse employee groups were more likely to engage in discussions about the effects of AI, not just on themselves but also on others in the organization. This led to a more cautious and restrained use of AI, illustrating *weak inseparability*. These findings demonstrate how the sociocultural AI literacy of humans influences ethical human-AI interactions.

In conclusion, **Paper V** provides empirical insights into how employees' critical and sociocultural AI literacy can strengthen their virtue ethics and, thereby, benefit ethical human-AI interaction (RQ1). AI-literate employees better assess and evaluate AI's outcomes and thus act more independently from AI's outputs, which can be explained by the sociomaterial notion of inseparability.

Final Conclusion for RQ1

To summarize the findings of Section 4.1, which address the research questions, "*How do human perceptions and the capabilities of both humans and AI influence ethical human-AI interaction?*," the paper's contributions are framed along the notions of sociomateriality, forming what is described as the sociomaterial cycle, as outlined in the theoretical background.

The findings reveal that the organizational environment—shaped by moral maxims, guidelines, and norms—affects human users' perceptions of what constitutes an ethical human-AI interaction (linked to the sociomaterial notion of *performativity*). In organizations emphasizing deontological principles, humans tend to perceive AI as rule-following technology, raising ethical concerns when AI systems make errors or decisions that deviate from societal norms (**Paper I**). When users lack education on AI's actual capabilities, they may overtrust AI and incorrectly attribute moral responsibility to AI, particularly when AI mimics human-like behaviors, thereby diminishing their own sense of virtue ethics by believing that the technology is reliable, safe, and aligned with human values (**Paper I**). Thus, human perceptions directly shape ethical judgments about the appropriate use of AI. These perceptions influence how humans interact with AI, underscoring the interdependencies between social and material entities that determine whether human-AI interactions are ethical (linked to the sociomaterial notion of *inseparability*). Within the interaction, AI's capabilities also impact how humans respond to it (**Papers I and III**). While AI can process large datasets and perform tasks with high precision, the way it executes decisions based on patterns may be invisible to human cognition and further lacks moral reasoning. Therefore, purely data-driven AI decisions may overlook important moral factors that a human decision-maker would naturally consider (**Paper V**). For instance, AI-literate individuals in security-oriented organizations are better equipped to critically assess AI's impact, detect biases, and raise ethical concerns, leading them to engage more cautiously and ethically (**Paper V**). Thus, the comparative capabilities of humans and AI systems introduce complex ethical dynamics that shape ethical human-AI interactions.

These dynamics further affect changes in human and AI agency (the sociomaterial notion of *relationality*). For example, overreliance on AI, possibly due to misconceptions about its capabilities (linked to *performativity*), can generate positive emotions that inhibit critical reflection on AI's outputs (**Paper III**), limiting human virtue ethics and moral agency and thus undermining ethical human-AI interactions. Conversely, when AI's operations lack transparency and users are overly dependent on its

outputs (linked to *inseparability*), this can trigger negative emotions and concerns about one's role (**Paper III**), prompting demands for greater oversight and explainability mechanisms to incorporate virtue ethics. These dynamics contribute to ethical human-AI interaction. Furthermore, the interaction is shaped by mechanisms, capabilities, and guidelines within the social subsystem, which influence both AI design and human behavior by establishing strategic directions, deontological principles, and utilitarian goals (linked to the sociomaterial notion of *materiality*). For instance, an organization's ability to balance exploration and exploitation strategies impacts its development of appropriate capabilities that can promote ethical AI design and deployment (**Paper IV**), contributing to ethical human-AI interactions. Or, the ethical design of AI involves considerations such as fairness, transparency, and inclusivity, ensuring that AI systems are developed and used in ways that respect human dignity and rights (**Papers I and II**). When an organization's and AI's capabilities are misaligned, and further contradict the workforce's moral values, this can result in misuse or biased algorithms, leading to morally problematic outcomes (**Paper I**) that hinder ethical human-AI interactions.

Finally, establishing routinized AI governance and responsibility practices is essential to ensuring that an organization's value systems align with its governance mechanisms (linked to the sociomaterial notion of *practices*, **Paper II**), thereby fostering ethical human-AI interactions. For instance, when an AI system makes a harmful decision, should the designers, users, or the system itself be held accountable? (**Paper I**) This question is further complicated by the perception of AI as either an autonomous agent (moral agent) or a mere tool (moral patient, **Paper I**), which varies widely depending on the specific context of interaction (linked to *performativity*). Ethical frameworks are needed to establish a sense of shared responsibility among all human stakeholders involved in AI design, deployment, and utilization and to clearly allocate responsibility between those human stakeholders (linked to *practices*; **Paper II**), especially when AI performs tasks traditionally executed by humans (linked to *inseparability*).

In conclusion, the ethical nature of human-AI interaction is shaped by the interplay between AI's capabilities (e.g., explanation mechanisms, autonomy), human capabilities—particularly AI literacy—and the capabilities of the organization's social subsystem to develop corresponding guidelines and align them with the workforce's value systems. The interplay between human perceptions, human capabilities, and AI capabilities can be illustrated through the sociomaterial cycle. When humans have a realistic understanding of AI and when deontological norms from the social subsystem align with human virtues and the AI's transparent design, ethical interactions are more likely. This alignment of AI design, human perceptions, and organizational guidelines fosters ethical behavior in human-AI engagements, ensuring that both human and AI capabilities contribute to responsible and transparent interactions.

4.2 Understanding Ethical Management of Human-AI Interactions

Regarding the second research question on how we can understand the ethical management of human-AI interaction, this dissertation includes three articles that establish theoretical foundations based on literature reviews (**Papers VI, VII, and VIII**) and two articles that provide empirical insights (**Papers IX and X**). Similar to Section 4.1, the sociomaterial notion that is most valuable for explaining each paper's contributions to RQ2 is used. While **Papers VII and IX** refer to the notion of practices, **Paper VIII** refers to the notion of materiality, and **Paper X** refers to the notion of relationality. **Paper VI**, as a foundational article, uses all notions of sociomateriality to illustrate their interconnection within the sociomaterial cycle, referring to the interplay of the different notions as illustrated in Section 2.2. In doing so, this section illustrates how the ethical management of human-AI interaction can be understood based on the interplay of several human-related and nonhuman objects (referring to the object level) within different work, organizational, and societal systems (referring to the system level) within the nested social subsystem, which shapes the human-AI interaction as sociomaterial entanglement.

Theoretical Foundations for RQ2: Paper VI (All Sociomaterial Notions in the Sociomaterial Cycle)

Paper VI: Heyder, T; Passlack, N. & Posegga, O. (2023) Ethical Management of Human-AI Interaction: A Theory Development Review. *The Journal of Strategic Information Systems (JSIS)*, 32(3), 101772, <https://doi.org/10.1016/j.jsis.2023.101772>.

The sixth paper of this dissertation addresses the emerging challenge of managing the interaction between humans and AI technologies, emphasizing the need for a comprehensive understanding of the ethical dimensions involved. It aims to ensure that AI deployment in organizations leads to positive outcomes, such as employee well-being, while avoiding unintended consequences like blurred lines of responsibility. The paper underscores the importance of integrating both duty ethics (deontology, focused on rules and principles) and virtue ethics (moral values) into the management of these interactions. Grounded in sociomateriality and sociotechnical system theory (e.g., Orlikowski, 2007, 2010; Orlikowski & Scott, 2008), the paper extends Leonardi's (2012) conceptualization of sociotechnical systems by incorporating literature on ethical management of human-AI interactions identified through a theoretical review. To offer a nuanced analysis, the paper applies five weak and strong notions of sociomateriality (inseparability, relationality, performativity, materiality, and practices; M. Jones, 2014) to explain the complex relationships between human virtue ethics, AI's duty ethics, and the ethical frameworks within the social subsystem of organizations. The paper employs a two-stage theoretical review approach. In the first conceptual stage, the paper connects constructs from sociotechnical system theory (as conceptualized by Leonardi (2012)) with ethical constructs—deontology and virtue ethics (e.g., Bowie, 1998; Eitel-Porter, 2021; Flathmann et al., 2021; Jobin et al., 2019; C. Jones et al., 2005)—as well as strategic management literature (e.g., Henderson & Venkatraman, 1999; Peppard & Ward, 2004). The second review stage systematically analyzes literature from seven databases, resulting in a final sample of 42 articles from 10,009 overall hits.

Paper VI's five key contributions to RQ2 are framed through the interplay of all sociomaterial notions, outlining their interdependencies within the sociomaterial cycle.

First, the paper contributes to understanding how inseparability can explain responsibility gaps in human-AI interactions. For instance, when AI systems act autonomously, humans may struggle to determine who holds responsibility for decisions or actions. In cases of *weak inseparability*, where humans and AI operate distinct tasks, humans may feel disempowered, particularly when AI operations lack transparency (i.e., when deontological principles are missing). This opacity creates moral buffers, distancing individuals from feeling accountable for the AI's outcomes, thus disconnecting them from their own virtue ethics. To address this, ethical management strategies require integrating transparent ethical principles into AI systems while also fostering human virtue ethics and a willingness to assume responsibilities through training and education. In contrast, *strong inseparability*, where human and AI tasks are interdependent (e.g., in algorithmic management through augmentation or assemblage), raises concerns about dehumanization, especially when AI's rigid deontological principles leave no room for the integration of human virtues. This rigid framework may lead to responsibility diffusion, where humans feel diminished moral agency as they merely follow AI directives without contributing to their own ethical decision-making. Therefore, ethical management requires aligning human virtue ethics with AI's deontological ethics to create a collective understanding of responsibility in the workplace. When this alignment is achieved (*strong inseparability*), it can positively influence the ethical design of AI (deontology) and shape humans' ethical perceptions of right and wrong (virtue ethics). This dynamic illustrates the connection between inseparability and relationality, contributing to the sociomaterial cycle, where the entanglement of humans and AI reshapes both technology and human agency.

Second, the paper outlines that different management approaches are required for *weak* or *strong relationality* in human-AI interactions, as this notion illustrates how the entanglement affects human agency, leading to either moral or immoral behavior. In cases of *weak relationality*, where human and AI capabilities remain distinct, immoral behavior can occur when humans intentionally misuse the AI system for their own benefit such as manipulating certain parameters to achieve better ratings in the context of crowdsourcing platforms. In such situations, the ethical management of AI must implement strict principles within AI systems to prevent misuse such as restrictions on AI's scope to ensure that humans cannot exploit the systems for unethical purposes. Additionally, the ethical management of humans requires establishing guidelines and mechanisms to ensure moral behavior such as bonus-malus systems, which judge human actions based on communicated ethical norms that exist independently of AI performance. This approach ensures that human actions are subject to ethical scrutiny regardless of AI's involvement, preserving individual moral responsibility. In the case of *strong relationality*, where AI alters human behavior and decisions more profoundly, ethical management must ensure that AI enhances rather than diminishes human moral agency. For instance, AI's neutral decision-making criteria in recruitment can guide human recruiters toward more objective evaluations, thereby enhancing

their moral behavior. However, in negative cases, AI's equal treatment may inadvertently lead to discrimination against disadvantaged groups. If human recruiters lack the ability to intervene, they might engage in unethical workarounds (e.g., loopholes) to mitigate such discrimination. Therefore, ethical management *in strong relationality* must address these interdependencies, ensuring that interaction with AI reinforces human moral agency rather than undermining it. This, in turn, can shape how individuals perceive and create their ethical reality, which is linked to the sociomaterial notion of performativity.

Third, the paper emphasizes that the notion of performativity is crucial for understanding how the beliefs, limitations, and boundaries of one's own virtue ethics—alongside those from the social subsystem and the AI artifact—shape human ethical perceptions and moral agency. For instance, leaders play a crucial role in shaping the ethical ideology of employees, influencing how they define and interpret ethics when interacting with AI. This demonstrates the *strong performativity* of the social subsystem on human moral agency, as the ethical behaviors modeled by leaders affect employees' understanding and treatment of AI technologies. However, this influence can also create ambivalence, as individuals may perceive AI differently across contexts. For example, employees may hold cautious, ethical views of AI in the workplace due to leaders' guidance but display more relaxed, careless attitudes toward AI at home. This reflects *weak performativity*, where external social beliefs do not fully influence human moral agency. Thus, ethical management must be aware that different cultural environments, with varied role models and contrasting societal values, can shape an individual's ethical perceptions of AI in diverse ways. This requires cultivating organizational norms and leadership practices that integrate human virtue ethics, ensuring that performativity strengthens rather than diminishes human moral agency in the workplace. By doing so, ethical management can establish practices within organizations (linked to the sociomaterial notion of practices) that shape how employees form their ethical realities (linked to performativity) within human-AI interaction, promoting a more ethically consistent interaction with AI across different contexts.

Fourth, the paper highlights the crucial role of the notion of practices in understanding the ethical management of human-AI interaction, particularly through the lens of how the entanglement of humans and AI impacts organizational practices. This entanglement, as seen in the notion of inseparability, shows that as human-AI interactions become more integrated, they can lead to significant changes in the way organizations conduct their operations. One key area of change in organizational practices is related to *strong practices*, where the entanglement of humans and AI evokes emotional responses from the workforce. For instance, when employees perceive moral violations during interactions with AI, they might resist further collaboration with AI systems if they feel that the AI's decisions or behaviors contradict their ethical values or result in unfair outcomes. This scenario represents *strong practices*, where the emotional response of the workforce necessitates a reevaluation of organizational practices, requiring managers to rethink how these interactions are managed to address the underlying ethical concerns. Additionally, the paper outlines how routinized practices emerge as a result of human-AI

entanglement. Over time, the consistent use of AI in everyday work can lead to the formalization of processes and norms within the organization. In a stronger form (*strong practices*), this could result in the creation of standardized work instructions or guidelines for how specific AI technologies should be used in decision-making or operational tasks. In the case of *weaker practices*, the sociomaterial entanglement may have little impact on existing work procedures, leaving organizational practices largely unchanged. Ethical management, therefore, must recognize the dual impact of human-AI entanglement on organizational practices, both in terms of affective responses and routinized behaviors. Leaders have to develop strategies that address the potential emotional concerns of employees and, similarly, establish formalized ethical guidelines for AI usage, which, in turn, impact how the AI artifact is arranged within the organization, illustrating a connection to the notion of materiality.

Fifth, the paper illustrates the importance of materiality in enhancing our understanding of how the management of different entities within a sociotechnical system can be combined to improve overall ethical human-AI interaction, given materiality's moderating role within the sociomaterial cycle, which impacts how all other notions can be aligned together. It emphasizes the role of transparent managerial strategies within all material entities that are somehow connected to the AI artifact and AI deployment decisions within organizations. Transparency is not limited to explaining the technical configuration of AI systems—such as their algorithms, data structures, and interfaces—but also extends to clearly communicating the organization's broader objectives regarding AI adoption. This transparency, as part of organizational practices (linked to the notion of practices), is essential for improving users' understanding of AI's actions and capabilities. It fosters more informed human-AI interactions by ensuring that individuals have a clear comprehension of how AI systems operate and what they aim to achieve. This helps to establish a realistic perception of AI (related to the notion of performativity) and strengthens the collaboration between humans and AI on equal terms (linked to the notion of inseparability). In doing so, transparency in AI's material configuration and its role within organizational practices enables more appropriate attribution of responsibility. When guidelines and systems are clearly communicated and understood, users are better equipped to assign responsibility where it belongs, ensuring that AI artifacts and the broader systems of which they are a part are ethically managed. This demonstrates a strong connection to the notion of relationality, as clear definitions of roles and responsibilities contribute to more ethical human-AI interactions. When the deontological principles guiding both AI and the material entities of the social subsystem surrounding it are aligned, this leads to more ethical outcomes in human-AI collaboration. Hence, *strong materiality* positively influences the relationship between the other four sociomaterial notions (inseparability, relationality, performativity, and practices) in what can be described as a sociomaterial cycle, contributing to a more ethical interaction between humans and AI based on deontology. Thus, ethical management must align the ethical principles of AI's technical configuration with the organization's practices to enable ethical human-AI interactions.

In conclusion, **Paper VI** advances our understanding of how to manage human-AI interaction ethically by demonstrating the usefulness of the sociomaterial notions in analyzing the ethical management of human-AI interactions (RQ2). It outlines how the interplay of sociomaterial notions can contribute to balancing the deontological principles of the AI with human virtue ethics within the sociomaterial cycle. In doing so, it contributes to the dissertation's theoretical background by integrating ethics into sociotechnical system theory, illustrating the value of the sociomaterial notions in explaining the dynamic relationships between duty and virtue ethics within the context of human-AI interactions.

Theoretical Foundations for RQ2: Paper VII (Practices)

Paper VII: Hammerschmidt, T. (2025). Navigating the Nexus of Ethical Standards and Moral Values. *Ethics and Information Technology (EIT)*, 27(17), <https://doi.org/10.1007/s10676-025-09826-5>.

The seventh manuscript of this dissertation aims to address the challenges in ensuring ethical human-AI interaction by navigating the interplay between the top-down implementation of ethical standards (linked to deontology) and the bottom-up integration of moral values (linked to virtue ethics) when aiming to manage human-AI interaction. It outlines the moral dilemmas that can occur when different stakeholders (e.g., developers, policymakers, managers, and users) hold varying normative views on ethics, resulting in different maxims linked to deontology, virtue ethics, or utilitarianism. The study outlines that different ontological frameworks, such as substantial and relational ontology, shape the ethical properties of entities within sociotechnical systems, building upon the work of Heyder et al. (2023) that extends Leonardi's (2012) conceptualizations to illustrate how different ethical views can clash. It employs a theoretical literature review of five databases with 2,161 overall hits and 26 articles in the final sample to explain how normative ethical theories can elucidate the dynamics between ethical standards and moral values in human-AI interaction, leading to moral dilemmas when they are not aligned.

Paper VII's three major contributions to RQ2 are framed by the notion of *weak* and *strong practices*, as it explains how to align different top-down and bottom-up managerial practices in organizations.

First, the paper highlights that for the ethical management of human-AI interaction, organizations must balance regulatory compliance (top-down) with moral deliberation (bottom-up). When different ethical views clash—such as AI adhering to utilitarian principles that prioritize the collective good, while individual users may have personal moral concerns—the organization must create platforms for ethical dialogue. This allows for the negotiation of ethical standards, facilitating moral sense-making among diverse stakeholders within an organization's environment. By doing so, organizations ensure that AI systems are not limited to following rule-based deontological principles, nor are AI strategies limited to the utilitarian goals of satisfying the needs of the majority of stakeholders; they also reflect individual human-centered values. This represents an example of *strong practices*, where ethical dialogue becomes a routine process for embedding virtue ethics into the design and implementation of

AI systems. Such practices ensure that AI not only performs tasks efficiently but also aligns with the moral values of its users, fostering a more ethical human-AI interaction.

Second, the paper offers pathways to resolve moral dilemmas by adopting context-specific approaches that account for varying ontological positions. These arise from different environments in which AI technologies interact with diverse human users, shaping perceptions of both human and AI agency. This variation requires distinct managerial practices to effectively balance both bottom-up and top-down approaches. For instance, in contexts where AI systems are used for crime detection, the ethical implications are closely tied to the institutions utilizing such algorithms for public safety. These systems influence institutional actions based on their outputs (linked to a relational ontology). The moral dilemma here lies in the utilitarian goal of maximizing public safety; however, it similarly raises concerns over data privacy and discrimination, as individuals not involved in criminal activities may be subject to surveillance and intrusive data collection, contradicting their virtue ethics. To address this dilemma, ethical management involves implementing several *strong practices* within the social subsystem to critically assess AI's outputs and encourage learning from public concerns regarding privacy and ethical treatment. The combination of these management practices might ensure that while AI-driven solutions contribute to public safety, they do not undermine fundamental ethical principles such as fairness and individual rights.

In contrast, in scenarios involving a substantial ontology, where human actions are less directly entangled with AI decisions, different moral dilemmas arise. For instance, in the context of autonomous vehicles, when these vehicles face unavoidable accidents in complex traffic situations with limited reaction time, the decisions made by the AI may have moral implications similar to instinctual human choices. However, the key ethical dilemma here concerns responsibility—who is accountable when an autonomous car's decision leads to harm? Unlike crime detection algorithms, where human agents are directly involved in interpreting and acting upon AI outputs, autonomous vehicles operate largely independently, which creates open questions regarding responsibility in accidents. As the AI systems in autonomous vehicles are tasked with making critical, life-or-death decisions, some may incorrectly attribute full moral agency to the AI itself. This misconception can result in human operators distancing themselves from responsibility, further complicating the ethical management of AI. Thus, ethical management must include regulatory frameworks as *strong practices* that determine the allocation of responsibility, enhance continuous human engagement, discuss trolley problems publicly, and integrate social welfare functions into the underlying models to ensure that while AI systems are autonomous, they are not seen as fully independent moral agents. Thus, the environment shapes different ontological positions that require context-specific approaches to organizational practices to contribute to the ethical management of human-AI interaction.

Third, the paper outlines the positive and negative influences of ethical standards on moral values and vice versa, thereby contributing to a more nuanced understanding of the ethical management of human-AI interaction. Ethical management should ideally foster a positive influence between ethical standards and moral values, ensuring that AI systems are aligned with societal expectations and human dignity. For instance, when AI designers consider community-based moral values when developing the technology by using participatory design approaches, these *strong practices* can ensure that the moral values of those involved in the AI design process positively contribute to broader ethical standards guiding AI development. Likewise, high ethical standards in AI can help prevent the dehumanization of individuals, ensuring that people are not treated merely as instruments to achieve particular goals, reflecting Kantian ethics. For instance, routines of having well-designed standards that give users guidance on ethical behavior (linked to *strong practices*), such as transparency protocol standards in AI, can mitigate potential moral hazards by fostering ethical reflection among users, helping them act more ethically in AI-driven interactions. However, negative influences can also arise. For example, when ethical standards are either too rigid or too narrowly focused, they may fail to account for the moral complexity of various contexts. Disconnections between formal legal requirements and the real-world moral responsibilities of users arise, illustrating an example of *weak practices* that hinder the ethical management of human-AI interaction. Hence, the dynamic relationship between moral values and ethical standards must be carefully considered when introducing managerial practices that aim to ensure ethical human-AI interaction.

In conclusion, **Paper VII** lays several theoretical foundations for our understanding of the ethical management of human-AI interaction by addressing how managerial practices can help handle moral dilemmas that arise when different ethical views clash in an organizational environment (RQ2). To better understand how to successfully deal with such potential moral dilemmas of conflicting moral maxims, the sociomaterial notion of practice is useful as it explains how to align affective practices that arise rather bottom-up from the workforce's virtue ethics with routinized practices of setting deontological principles and utilitarian goals rather top-down through the regulation of politicians' or leaders' strategy-making. Thus, the article lays the theoretical foundation that ethical management of human-AI interaction requires managerial practices that respect diverse ethical perspectives and allow for human moral agency in decision-making (linked to the notion of practices).

Theoretical Foundations for RQ2: Paper VIII (Materiality)

Paper VIII: Hammerschmidt, T.; Stolz, K. & Posegga, O. (2025). Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI. Behaviour of Information Technology (BIT), <https://doi.org/10.1080/0144929X.2025.2500451>.

The eighth paper sheds light on how the absence of certain skills, capabilities, and resources, or misleading perceptions, shapes ethical outcomes in human-AI interactions, particularly in relation to managing the creation of sustainable outcomes. It aims to increase our understanding of how inequalities

in access and skills limit the potential for AI to create and achieve sustainability goals. As AI becomes increasingly embedded in societal and organizational systems, the gap between those who can and cannot harness AI's benefits—referred to as the “AI divide”—is widening (Carter et al., 2020; Klumpp & Zijm, 2019; Vassilakopoulou & Hustad, 2023). While human-AI interactions can result in various outcomes, such as increased efficiency and improved performance, this study emphasizes the need to prioritize sustainable outcomes when managing human-AI interaction, which considers the social, ethical, and environmental impacts of AI usage linked to the United Nations' Sustainable Development Goals (SDGs), particularly SDG#10 (reducing inequality), SDG#4 (quality education), and SDG#16 (inclusive societies) (Fosso Wamba et al., 2024; Goralski & Tan, 2020). Sustainable outcomes can be seen as an integral part of managing ethical human-AI interaction, particularly when they focus on promoting equity, inclusivity, and long-term well-being. Sustainability and ethics are linked in this context because both involve the fair distribution of resources and opportunities across present and future generations, ensuring that the benefits of AI technologies are accessible to all without causing harm to people or the planet. The study builds upon the digital divide literature, addressing three levels of inequalities: the access divide (first-level divide), which refers to unequal access to digital technologies; the usage or skill divide (second-level divide), concerning disparities in the ability to use digital technologies effectively; and the outcome divide (third-level divide), which highlights the inequalities in achieving valuable outcomes with these technologies (Riggins & Dewan, 2005; Van Deursen & Van Dijk, 2011; Wei et al., 2011). Using sociotechnical system theory as outlined by Leonardi (2012), the paper contextualizes different social, technical, and sociotechnical drivers of AI inequalities in alignment with Carter et al. (2020). Through a theoretical literature review of 8,709 initial hits and a final sample of 59 articles, the study employs a keyword search across seven databases to conceptualize the AI divide from a sustainable development perspective, grounded in sociotechnical systems theory.

Paper VIII's three major contributions to RQ2 are framed by the notion of *weak* and *strong materiality*, as it explains how the technological artifact and other nonhuman objects in an organization's social subsystem must be aligned to ethically manage human-AI interactions.

First, the study emphasizes the importance of recognizing contradictory effects between the divide levels when managing human-AI interaction ethically. For instance, while the increased application of AI in organizations can enhance task efficiency and productivity, thereby reducing the AI outcome divide, its rapid and unchecked introduction into the workforce can lead to unintended negative consequences such as employees struggling with their roles. When AI is deployed extensively (reflecting *strong materiality* in terms of technological integration) but is not supported by appropriate communication strategies or employee engagement efforts (reflecting *weak materiality* within the social subsystem), this can diminish the workforce's sense of purpose, widening the outcome divide—not in terms of productivity but in terms of employee satisfaction and workplace well-being. Thus, the paper

outlines that it is essential for managerial strategies to address not only the technical aspects of AI deployment but also the human and social dimensions that contribute to ethical human-AI interactions. *Strong materiality* between both the material artifact and the social subsystem is required to ensure that the benefits of AI—such as increased productivity—do not come at the cost of employee well-being or organizational cohesion. This illustrates a utilitarian approach, aiming to maximize AI benefits for most employees. *Strong materiality* ensures that AI technologies are not only efficiently integrated but are also aligned with the broader social and ethical goals of the organization, fostering a sustainable approach to ethically managing human-AI interaction.

Second, the article highlights the importance of considering the interdependencies between the properties of AI as a technological artifact and the social subsystem when aiming to foster ethical human-AI interaction. Due to the circular nature of the AI divide levels, the interaction of these properties can either mitigate or exacerbate inequalities, thereby influencing the creation of sustainable or unsustainable outcomes. This dynamic can be better understood through the sociomaterial notion of materiality, which describes how the material aspects of the technology are organized, designed, and arranged within other material aspects of the social subsystem. For instance, human-like capabilities and the anthropomorphic design of certain AI technologies such as smart voice assistants, can reduce access barriers by making the technology more intuitive and relatable to users. This encourages broader adoption and increased usage as individuals find it easier to engage with these systems. As users gain more experience with these technologies, the initial reduction in the access divide leads to a corresponding decrease in the usage divide. In this scenario, we see the manifestation of *strong materiality*, where the human-like behavior of the AI is directly tied to its anthropomorphic design. However, without proper guidance or transparent communication from the social subsystem (e.g., service providers or regulatory bodies), the anthropomorphic design of the AI could obscure the inherent biases or limitations of the technology. In this case, the failure of the social subsystem to provide adequate oversight or education weakens the intended reduction of access and usage barriers. This reflects *weak materiality*, where the material components of the AI design of the artifact and the AI guidelines from the social subsystem are not effectively aligned. Thus, both the technological artifact and the social subsystem must exhibit *strong materiality*. If one aspect is weak while the other is strong, the reduction of inequalities at one divide level may lead to an increase in another. Consequently, this work highlights the critical interdependencies between the properties of the technological artifact and the social subsystem in shaping the reduction or increase of AI inequalities across different divide levels.

Third, the paper emphasizes the role of aligning certain properties (e.g., resources, guidelines, strategies) of the nested social subsystems, comprising various stakeholders from multiple layers, in shaping ethical human-AI interactions. While top-down regulations, such as the European AI Act from the societal layer, provide overarching governance mechanisms for industries, organizations, and development teams, there are also bottom-up initiatives from developer communities that help shape

future AI regulations. To achieve a balanced alignment between top-down and bottom-up regulatory approaches within these nested layers, the social subsystem must exhibit either *weak* or *strong materiality* to equally contribute to creating sustainable outcomes in human-AI interactions and ultimately reducing the outcome divide. For instance, in the realm of AI governance, the establishment of strict regulations and robust governance frameworks at the societal layer (reflecting *strong materiality*) can provide essential guidance for ethical AI design and deployment. However, if these regulations become overly restrictive and fail to consider the practical needs and legal requirements of developer teams or AI service providers, they may hinder innovation. This misalignment can create barriers to norm adherence, discouraging AI service providers from developing technologies that contribute to social good, which exemplifies *weak materiality*. Another example can be seen in AI strategies pursued by governments aiming to boost AI deployment (*strong materiality*), which may result in regional inequalities, such as widening the productivity-pay gap between urban and rural areas. To avoid further inequalities in regional development that contradict utilitarian goals, corresponding efforts from regional political initiatives and organizational strategies that account for these disparities are required (which would also represent *strong materiality* as a required counterpart). This illustrates that material interdependencies between various layers within the social subsystem must be aligned to reduce the outcome divide and, thus, foster ethical management of human-AI interactions.

In conclusion, **Paper VIII** advances our understanding of how to manage human-AI interaction ethically by examining the *weak* and *strong material* interdependencies that lead to inequalities across different divides. These interdependencies either hinder or promote the creation of sustainable outcomes through human-AI interactions. The paper provides a theoretical foundation for addressing RQ2 by demonstrating the relevance of the sociotechnical system theory and the notion of materiality to illustrate the interconnected management mechanisms that can either strengthen or weaken inequalities, thereby influencing the ethical management of human-AI interactions by following utilitarian goals.

Empirical Evidence for RQ2: Paper IX (Practices)

Paper IX: Hammerschmidt, T.; Passlack, N. & Posegga, O. (working paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment. [*previously 2nd revision round at JSIS*].

The ninth paper of this dissertation addresses the challenge of ethically managing the design, deployment, and utilization of AI in organizations, highlighting several decision dilemmas that leaders face when determining how to integrate AI to generate business value. The theoretical background builds on the work of Heyder et al. (2023), integrating different ethical perspectives into Leonardi's (2012) conceptualization of sociotechnical systems. Additionally, the paper draws on ambidexterity literature to emphasize the importance of balancing exploration and exploitation in decision-making (e.g., Mom et al., 2019; O'Reilly & Tushman, 2004, 2011). The paper outlines that focusing too heavily on exploration can result in high ethical risks, while an overemphasis on exploitation may lead to missed

AI benefits. The study employs semi-structured interviews with 27 German-speaking experts, including leaders responsible for AI deployment within organizations and employees who adopt AI unsupervised in their daily tasks, to explore these dilemmas and the ethical considerations surrounding AI integration from different perspectives.

Paper IX's two key contributions to RQ2 are framed by the notion of weak and strong practices, as it explains how to integrate ethical human-AI interaction routines and procedures within the organization's social subsystem.

First, the findings highlight the importance of incorporating both shared responsibilities and clear responsibility allocations into managerial practices for ensuring ethical human-AI interaction, especially when multiple stakeholders (e.g., leaders, users, and designers) are involved in AI-related activities. Interviewees identified a dilemma between fostering a culture of shared responsibility and establishing clear responsibility allocation when integrating new AI technologies for business growth (linked to exploration strategies) within organizations. While interviewees acknowledge that a shared responsibility culture can promote collective ownership, it also risks individuals attributing responsibility solely to the AI or shifting blame to others, particularly when dealing with opaque AI systems, where the decision-making processes are not transparent. To address this dilemma, interviewees suggested combining *strong practices* that clearly define responsibilities through guidelines and duties (aligned with deontological principles) with practices that foster a culture of shared responsibility, encouraging stakeholders to take accountability based on their virtue ethics. This approach cultivates moral responsibility and ethical behavior, ensuring that all stakeholders actively engage responsibly with the consequences and implications of AI when pursuing exploration strategies to leverage AI innovations for business value creation.

Second, the findings further emphasize balanced approaches between process standardization through AI and allowing for more individualized human decision-making when managing human-AI interactions ethically. Interviewees noted the dilemma they face when integrating AI to optimize existing procedures (linked to exploitation strategies), questioning whether to use AI's neutral and objective logic to standardize operations or to preserve space for human intention, judgment, and creativity. Standardization through AI, grounded in deontological principles, can enhance efficiency and objectivity (e.g., in recruitment or performance evaluations), but it may also lead to the loss of human uniqueness and personalized attention to individual circumstances, thereby diminishing the role of human virtue ethics. To address this dilemma, the ethical management of human-AI interaction requires a combination of *strong practices*: the establishment of routinized, AI-driven procedures to create a consistent baseline of ethical standards that will fit the majority of organizational stakeholders (linked to utilitarianism), alongside the inclusion of critical decision points where human reflection and input are integrated (linked to virtue ethics). This balanced approach ensures that while AI enhances

operational efficiency, human judgment can still be incorporated to address cases requiring personalized attention, maintaining an ethical focus on individual needs within managerial practices.

In conclusion, **Paper IX** offers valuable empirical insights into understanding complex decision dilemmas that arise when organizations integrate AI technologies into existing processes to optimize working efficiency (linked to exploitation) or use AI innovations for business growth (linked to exploration). It offers various pathways (linked to the sociomaterial notions of practices) to address those decision dilemmas, thereby providing ideas on how we can understand the ethical management of human-AI interaction (RQ2).

Empirical Evidence for RQ2: Paper X (Relationality)

Paper X: Stolz, K.; Hammerschmidt T. & Posegga, O. (2024). How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making (Best Paper Nominee). Proceedings of 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, <https://hdl.handle.net/10125/106970>.

The tenth paper of this dissertation focuses on the challenge of promoting sustainable consumption, especially in the fashion industry, which faces significant environmental impacts. It explores the use of nudging mechanisms through conversational agents to encourage consumers toward environmentally friendly shopping decisions, thereby contributing to increased ethical behavior in human-AI interaction. This work builds on the nudging categories of Hansen and Jespersen (2013), as adapted by Caraban et al. (2019), differentiating between transparent and non-transparent nudges, as well as between nudges that promote reflective thinking compared to those that encourage automatic thinking. Additionally, the work refers to green IS research (e.g., Alt, 2020; Corbett, 2013), focusing on how technologically mediated processes can foster green decision-making, such as promoting more sustainable product alternatives. The exploratory study conducted a qualitative online survey, collecting responses from 79 fashion shoppers to derive propositions regarding users' perceptions of conversational agents attempting to influence their shopping behavior.

Thus, **Paper X's** two major contributions to RQ2 are framed by the notions of *weak* and *strong relationality*, as it explains how human-AI interaction can alter aspects of human agency (e.g., their behavior) as an outcome of sociomaterial entanglement.

First, the paper highlights that implementing nudges within AI artifacts such as conversational agents has the potential to influence consumer shopping behavior by encouraging the selection of more sustainable fashion products. Participants indicated that certain nudges—such as presenting fair-trade labels or comparing the carbon footprint of different products—made sustainable alternatives more visible and accessible. In these cases, participants who identified as sustainability-oriented reported that they would follow these nudges, resulting in more sustainable purchasing decisions. This scenario demonstrates *weak relationality*, as the nudge merely facilitates ethical choices that align with

preexisting values, reducing the effort required to find sustainable products. Here, the ethical management of human-AI interaction involves supporting individuals' preexisting ethical intentions. Conversely, participants who did not prioritize environmental concerns during online shopping required more direct or confrontational nudges, which prompted them to reflect on their purchasing behavior. In such instances, the AI's capacity to adapt to user preferences and deliver personalized recommendations enables more nuanced and effective nudges that can alter consumer behavior, exemplifying *strong relationality*. This relationality reflects a deeper interaction between the AI and the user, where the AI actively encourages behavior change rather than simply supporting existing tendencies. However, when nudges are not tailored to individual preferences, they may be perceived as intrusive or disruptive, leading users to disregard the agent's prompts—another example of *weak relationality*. This finding underlines the importance of personalization in ensuring that nudges are not only effective but also aligned with users' values and preferences. Thus, the ethical management of human-AI interaction, particularly in the context of promoting sustainable shopping behavior, involves creating a balance between supporting users' ethical intentions and encouraging reflection for those who may not initially prioritize ethical maxims such as sustainability goals. This balance can be achieved through the strategic use of personalized nudges that foster both *weak* and *strong relationality*, depending on the user's needs and values.

Second, the paper emphasizes the importance of designing ethically compliant technological artifacts that are transparent and aligned with values such as security, safety, accountability, and inclusivity. The study reveals that participants' perceptions of adherence to ethical standards are enhanced when AI systems are intentionally designed to achieve ethics-related goals. While participants generally supported the use of nudges within conversational agents to promote sustainable consumption, they also expressed that sustainability-related recommendations entail higher expectations for ethical compliance compared to other AI applications. This represents *strong relationality*, as the use of nudging mechanisms within human-AI interaction for ethical maxims results in higher expectations of AI's ethical compliance. Participants argued that the use of nudges, as well as the objectives they seek to achieve, should be clearly communicated in a transparent manner. Concerns were raised about issues of transparency, fairness, and the risk of manipulation, which participants considered central to the ethical management of AI systems. If conversational agents are perceived as ethical and trustworthy, users are more likely to accept the nudging mechanisms employed to influence their decisions (*strong relationality*). If nudges are subtle and not clearly communicated, users may not recognize them as intentional prompts, and they remain unaware of the influence these nudges exert on their decision-making (*weak relationality*). Thus, the acceptance of nudges that guide users toward ethical maxims—whether through weak or strong relationality—ultimately depends on the ethical design of the AI system. In conclusion, ethically compliant design is crucial not only for fostering trust but also for ensuring the long-term effectiveness of AI-driven nudging mechanisms in encouraging sustainable and ethical behavior.

In conclusion, **Paper X** provides empirical insights into how the design mechanisms of conversational agents can facilitate ethical human-AI interactions, particularly in promoting sustainable consumption. By differentiating between *weak* and *strong relationality*, the findings illustrate that the ethical management of AI requires a careful balance between supporting users' preexisting ethical intentions and encouraging reflection among those who may not prioritize sustainability. Linked to RQ2, the study underscores the necessity of transparent and ethically compliant management mechanisms (e.g., ethical AI design) when aiming to encourage more ethical behavior such as more sustainable consumption decisions.

Final Conclusion for RQ2

To summarize the findings of Section 4.2, addressing the research question “*How can we understand ethical management of human-AI interaction?*,” the paper's contributions can be framed along the notions of sociomateriality, forming the sociomaterial cycle as a major contribution of **Paper VI**. This cycle highlights how ethical management strategies and practices need to be aligned with the diverse ethical perspectives that appear within human and AI technologies, as well as the social subsystem. The findings demonstrate that the ethical management of human-AI interaction can be understood as a dynamic interplay between sociomaterial notions, which collectively form a cycle of managerial strategies and practices designed to ensure ethical behavior and outcomes when humans interact with AI, thus fostering ethical human-AI interaction in organizations.

To promote realistic human perceptions of AI capabilities and limitations, ethical management must shape users' ethical perceptions by raising awareness of the fundamental differences between human and AI moral agency (linked to the sociomaterial notion of *performativity*). This involves managerial practices that (among others) establish leaders as role models who exemplify ethical behavior, alongside cultivating a corporate culture that is sensitive to the nuances of human-AI interaction (**Paper VI**). For example, training programs, ethical guidelines, and communication strategies that clarify AI's role are essential for preventing misattributions of moral responsibility to AI technology (**Paper VI, VII**). This further impacts the interdependencies between human and AI agency within the interaction (linked to the sociomaterial notion of *inseparability*). To prevent AI from displacing or undermining human agency within interactions, ethical management must ensure that there is space for human oversight, interpretation, and intervention in critical decision-making processes (**Papers VI and IX**). This is particularly important in situations where AI handles complex tasks beyond human capacity such as processing large volumes of data, but human moral judgment remains essential to ensure ethical outcomes (**Paper VII**). Therefore, managing human-AI interaction ethically requires ensuring that users are trained to interact with AI technologies ethically, making them AI-literate in understanding their benefits and risks (**Paper VI**). Ethical management must further ensure a balance between data-driven efficiency and human virtue ethics, as both create a symbiotic relationship in which AI contributes to

operational tasks while humans provide ethical oversight and moral discernment (**Papers VI and VII**). Ethical human-AI interaction, therefore, requires that humans retain control over the final decision in high-stakes environments, such as autonomous vehicles, medical diagnostics, or criminal justice, to take into account broader ethical and social considerations that AI systems may overlook (**Paper VII**).

Furthermore, the management strategies and practices involved in shaping ethical human-AI interaction (linked to *inseparability*) should further be aligned with those designed to influence human behavior within the interaction (linked to the sociomaterial notion of *relationality*). For instance, AI nudging mechanisms can positively influence human behavior by encouraging sustainable and ethical decision-making (**Papers VI, IX, and X**). However, it is essential that these nudges align with individuals' existing ethical intentions, reinforcing them rather than creating contradictions or ethical conflicts (**Papers VI, VII, and X**). Thus, ethical management should ensure that AI nudges support reflection, rather than undermining or manipulating users. Therefore, ethical management of human-AI interaction requires alignment between the AI artifact's design and the corresponding norms, guidelines, and policies within the organization's social subsystem (linked to the sociomaterial notion of *materiality*). Organizations must be attuned to the interplay between social (human-related), technical (AI-related), and sociotechnical factors that influence the ethical outcomes of human-AI interactions (**Paper VIII**). For instance, the careful alignment of AI design features with organizational goals and social values can help mitigate inequalities across different levels of the AI divide, preventing the exacerbation of issues such as bias or unequal access to technology (**Paper VIII**). Ethical management, therefore, should drive public consultations on how AI should be developed and integrated in organizations ethically by engaging in stakeholder dialogues—including industry leaders, ethicists, policymakers, and civil society (**Papers VII and VIII**).

Finally, the ethical management of human-AI interaction must also account for potential moral conflicts that arise when diverse ethical perspectives—such as utilitarianism, deontology, and virtue ethics—clash within human-AI interactions (**Paper IX**). This necessitates the establishment of practices that not only routinize ethical behavior but also manage the emotional and ethical responses of the workforce to AI technologies (linked to the sociomaterial notion of *practices*). Organizations have to foster ethical resilience and routines to increase awareness of potentially conflicting moral maxims (related to different ethical perspectives) and to integrate those different perspectives within managerial strategies and practices for AI design, deployment, and utilization (**Papers VI, VII, and IX**). This involves balancing top-down regulation (involving internal guidelines and external regulations with ethical principles such as beneficence, transparency, and accountability) within the nested social subsystem and bottom-up ethical initiatives and moral deliberation within the workforce to ensure that rule-based deontological principles of AI design, utilitarian maxims of AI deployment strategies, and individual human values are reflected (**Papers VII and IX**). Thus, managerial practices for ethical human-AI interaction involve ensuring that developers and AI providers are responsible for the ethical

design of AI systems, users understand their role in overseeing AI systems, and organizations take corrective actions within their AI deployment strategies when ethical issues arise (**Papers VI, VII, and IX**).

In conclusion, the ethical management of human-AI interaction requires multifaceted approaches that integrate and align different managerial strategies and processes with ethical frameworks, as illustrated through the sociomaterial cycle. Therefore, ethical management must go beyond merely responding to the challenges and dynamics inherent in human-AI collaboration. Instead of being reactive when aiming to manage human-AI interaction ethically—addressing issues only as they arise from the evolving interplay of human and AI capabilities (as outlined in Section 4.1)—ethical management should be proactive. This means shaping an environment where ethical considerations are embedded from the outset, driving the development and application of AI technologies in a way that anticipates potential ethical dilemmas and fosters responsible outcomes. This requires continuously refining governance structures, ethical frameworks, and oversight mechanisms to stay ahead of technological advancements, ensuring that AI development remains aligned with societal moral values and human rights.

5 Discussion

This section outlines the theoretical (Section 5.1), practical (Section 5.2), and methodological (Section 5.3) implications of this dissertation. Additionally, it outlines the limitations and provides suggestions for future work (Section 5.4).

5.1 Theoretical Implications

This dissertation makes theoretical contributions to IS research by advancing our understanding of ethics within human-AI interactions (addressing RQ1) and of ethically managing these interactions (addressing RQ2). It builds on the premise that ontological analysis—whether framed through relational or substantial ontology—must precede any ethical inquiry (Hammerschmidt, 2025). Such an analysis involves examining the nature and relationships of social and material entities within sociotechnical systems to better understand how ethical perspectives are integrated into human-AI interactions. This approach allows for a thorough explanation of how the capabilities of human, AI, and the organization's social subsystems form ethical human-AI interactions (Section 4.1), leading to the development of strategies for their ethical management (Section 4.2). Thus, this dissertation's theoretical implications are twofold: *first*, the dissertation contributes to the discourse on sociomateriality and moral philosophy by demonstrating how the lens of sociomateriality can be employed to outline the non-neutral co-constitution of human and AI moral agency, which differs depending on ontological positions and represents an antecedent to ethical human-AI interaction; *second*, it contributes to research on the ethical management of human-AI interaction by revealing the boundaries that arise when connecting ethical theories with sociomaterial discourse.

Ontological Positions Shaping the Co-constitution of Human and AI Moral Agency: Antecedents to Ethical Human-AI Interaction

This dissertation highlights the relevance of addressing the foundational ontological questions on how we understand AI's moral agency in relation to human moral agency, before applying ethical theories (like deontology, virtue ethics, and consequentialism) to human-AI interactions, as this understanding shapes how ethical concepts are framed and applied. Thus, before understanding ethics within human-AI interactions, we have to understand existing views on AI and human moral agency. Sociomateriality, which closely aligns with relational ontology in a strong form and substantial ontology in a weak form, provides a valuable lens for illustrating the non-neutral co-constitution of human and AI moral agency. The term "co-constitution," as employed in this dissertation, draws on Leonardi's (2011, p. 151) "metaphor of imbrication," which describes the varying ways in which human and material agency can be interwoven. In a weak form of imbrication (corresponding to *weak sociomateriality*), human and AI moral agencies remain independent and stable throughout their interaction, reflecting the substantial ontology perspective (Hammerschmidt, 2025; Heyder et al., 2023). By contrast, in a strong form of imbrication (corresponding to *strong sociomateriality*), human and AI moral agencies are inseparable and co-constructed through their interactions, aligning with relational ontology (Hammerschmidt, 2025; Heyder et al., 2023). The non-neutrality stems from the fact that AI systems are not merely passive technologies driven by human intent (as moral patients); rather, they are increasingly becoming moral agents that can influence and shape moral outcomes (Hammerschmidt, Hafner, et al., 2025). Thus, both ontological positions are crucial to consider when aiming to understand ethical human-AI interactions because they enable us to comprehend the implications of human-AI interactions for both human and AI moral agency, as illustrated in Figure 2:

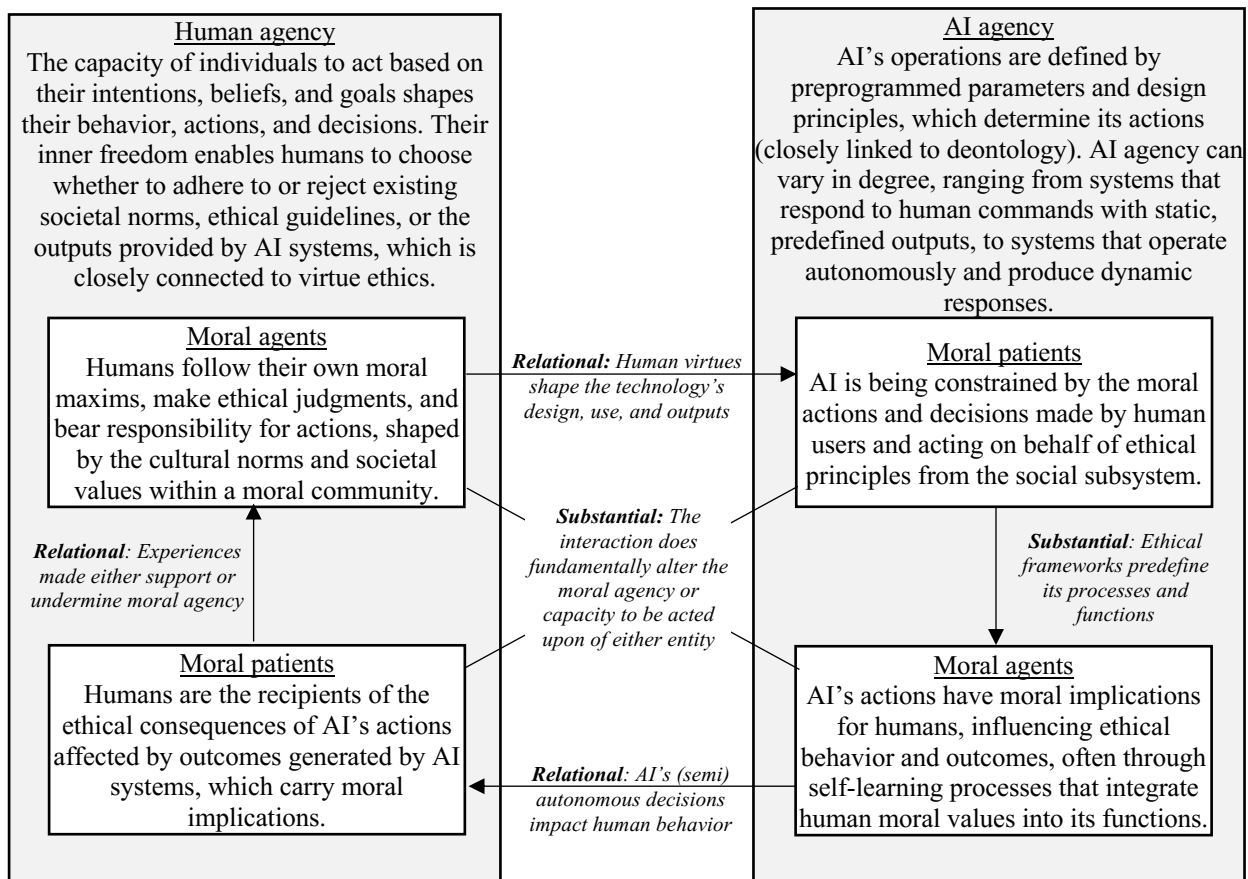


Figure 2: Ontological positions shaping our understanding of the co-constitution of human and AI moral agency.

Figure 2 presents an ontological framework that provides a nuanced understanding of the different ontological perspectives of the co-constitution of human and AI moral agency. Within this framework, both humans and AI occupy dual roles as moral agents and moral patients, each contributing to or hindering ethical human-AI interactions. Human agency is closely aligned with virtue ethics, focusing on individual moral character, while AI agency is generally linked to deontological principles, as its actions are often based ethical principles. The arrows depicting relational and substantial dynamics underscore the importance of both the structural and behavioral aspects of human-AI interactions. Relational aspects highlight the ongoing influence between entities, while substantial aspects show how behavior, decisions, and actions remain stable, irrespective of concrete interactions.

As depicted in Figure 2, within a relational ontology, the ethicality of the interaction is shaped by a dynamic interplay between the two dual roles of moral agents and moral patients. In their role as moral agents, humans exercise their moral maxims and ethical judgment abilities to shape the design, implementation, and functioning of an AI system (**Paper I** and **VII**). For instance, developers integrating nudges in the design of conversational agents to drive sustainable decisions (**Paper X**), dictating how the AI system functions as moral patient, and thus, reflecting the moral values of their human creators. On the other hand, as moral agents, AI technologies can influence the behavior of decision-making processes of humans, for instance, when those AI nudges work and users change their decision behavior

(**Paper X**), making them to moral patients. In this capacity, humans are moral patients, experiencing the ethical consequences of AI's nudges, in turn impact human's own moral judgments and actions in future interactions with AI, thus, impacting their moral agency. A critical ethical consideration here is the need for alignment between human moral maxims and AI operations to prevent moral relativism when the relational dynamics within the interaction allow for justifying any action based on selective moral perspectives (see **Papers I** and **VII**). Without such alignment (*weak sociomateriality*), AI systems could act in ways that contradict or undermine human moral values (see **Papers VII** and **IX**); however, within too strong relations (*strong sociomateriality*), it risks ethical inconsistency in how ethical reality is created between both entities in the interaction and the reality of AI regulation and laws (see **Papers I** and **VII**).

Conversely, within a substantial ontology, human and AI moral agencies remain stable and do not undergo fundamental shifts during their interaction, as their properties are viewed as distinct and unchanging. In this view, AI's ethical principles and operational mechanisms—though initially shaped by human designers—exist alongside the specific user's ethical intentions. Thus, human's as moral agents does not alter the moral agency of AI, and vice versa, AI's moral agency does not fundamentally impact human moral values and virtues. Therefore, the ethical frameworks on which AI technologies are programmed on as moral patients also determine their actions as moral agents. This separation is particularly crucial when users have malicious intentions or seek to exploit AI systems for unethical purposes (see **Paper VI**). In such cases, the AI's ethical principles must remain distinct and robust enough to prevent misuse. Moreover, maintaining a separation between AI's neutral operational principles and human behavior is important to avoid excessive alignment that could lead to a lack of humanity and empathy. For example, in decision-making processes where AI is involved, such as criminal sentencing or medical diagnostics, overreliance on AI's neutral, data-driven conclusions without considering human empathy could result in ethically problematic outcomes (**Paper VII**). Therefore, the key ethical concern is twofold: ensuring that AI technologies are developed with robust, transparent ethical guidelines, and simultaneously strengthening human moral maxims and ethical behavior to ensure that users are not negatively influenced by their interaction with AI technologies.

To summarize, neither relational nor substantial ontology, by itself, provides a complete framework for understanding ethical human-AI interaction. Instead, this dissertation argues for the necessity of combining both positions to capture the complexity of these interactions. A dual approach allows for a more comprehensive understanding of how AI systems—whose agency can evolve over time—interact with humans, whose moral framework might shift depending on various contextual factors. By considering both ontological views, we can develop a more nuanced understanding of how to manage ethical human-AI interactions as before understanding the ethical “right” and “wrong” of human-AI interactions (linked to ethical agency), it is crucial to first understand the factors shaping “good” and “bad” behavior within these interactions (linked to moral agency); (**Paper I**). Thus, the analysis of how

ontological positions shape our understanding of human and AI moral agency act as key antecedents for understanding ethical human-AI interactions. This dissertation, therefore, offers new insights into how the integration of both ontological position through the weak and strong notions of sociomateriality can deepen our understanding of human and AI moral agency within human-AI interaction. It addresses the gaps illustrated by Cecez-Kecmanovic (2014) and Jones (2014), who called for a more nuanced integration of critical (relational ontology) and agential (substantial ontology) realist perspectives into the sociomateriality discourse. In doing so, this work provides a more holistic framework for understanding the ethical management of human-AI interactions.

Ontological Boundaries of Combining Sociomateriality with Ethical Theories

This dissertation further highlights that traditional ethical theories, by nature, align more closely with a substantial ontology, while the emerging challenges of human-AI interactions point toward a relational ontology (for instance, when comparing the insights of **Paper I to IX**). This distinction introduces certain boundaries when interpreting ethics in the context of human-AI interaction. For example, virtue ethics, rooted in Aristotelian philosophy, presupposes that humans possess inherent qualities (virtues) that define their moral character (e.g., Hammerschmidt, Hafner, et al., 2025; Heyder et al., 2023), grounded in a substantial ontology. Similarly, consequentialism evaluates the ethical value of actions based on outcomes, assuming that entities (whether human or AI) possess stable properties, such as rationality, that allow them to evaluate actions based on potential benefits or harms. Kant's categorical imperative, understood as a universal law, further exemplifies a deontological approach where ethical principles are fixed and unchanging (e.g., Hafner et al., 2024; Hammerschmidt, Hafner, et al., 2025; Heyder et al., 2023), aligned with substantial ontology. Accordingly, ethical management focuses on entities possessing stable and intrinsic properties, whether that involves cultivating virtues in human developers or users or improving AI rationality within a consequentialist framework.

In contrast, a relational ontology posits that entities are not defined solely by their intrinsic properties but by the relationships and interactions in which they engage. From this perspective, ethical obligations and responsibilities emerge from the interactions between entities, such as humans and AI, highlighting that moral agency is co-constituted through these relationships. Rather than focusing exclusively on cultivating human virtues or improving AI's decision-making outcomes, this approach considers how ethical responsibilities arise from the dynamic interplay between human and AI entities. This dissertation, therefore, does not advocate replacing substantial ontology with a relational approach; instead, it shows how traditionally ontological theories of ethics can be expanded through an understanding of the non-neutral, co-constitutive nature of human-AI interactions (e.g., Hammerschmidt, 2025; Heyder et al., 2023).

However, this relational perspective has its boundaries. For instance, the dissertation often refers to the idea of aligning AI's deontological principles with human virtue ethics—an alignment that reflects a relational ontological stance (e.g., Hammerschmidt, 2025; Hammerschmidt, Passlack, et al., submitted). Yet it is important to recognize that the intrinsic properties of humans and the stable characteristics of technological artifacts can only be aligned to a certain extent. This tension could, in some cases, undermine efforts to cultivate human virtues. For instance, in AI-driven hiring processes, if the system is designed to follow strict deontological principles, recruiters might feel compelled to prioritize AI's recommendations over their own ethical judgments, reducing empathy in decision-making (e.g., Hammerschmidt, 2025; Heyder et al., 2023). Thus, human virtues might be reduced to mere compliance with AI outputs, limiting the sensibility for ethical reasoning that recruiters usually employ (e.g., Hammerschmidt, 2025; Heyder et al., 2023). Hence, aligning humans too closely with AI systems might reduce human virtues to mere alignment goals, which could be counterproductive to the broader aims of establishing ethical human-AI interactions.

Table 4 illustrates the intersections and boundaries between normative ethical theories and the two different ontological positions that are embedded in this dissertation's sociomaterial analysis of ethical human-AI interactions.

Ontology	Normative Ethical Theory		
	Deontology	Virtue Ethics	Consequentialism
Relational Ontology (<i>Strong Sociomateriality</i>)	• Ethical human-AI interactions mean that ethical principles evolve dynamically during interactions between humans and AI. • Ethical management requires adjusting the ethical duties of both AI and humans according to the specific social context and situational demands of the application. • Boundaries: The dynamic nature of evolving ethical principles can lead to ethical instability and inconsistencies across different contexts.	• Ethical human-AI interaction involves AI learning from human data input, integrating human virtues (to a limited extent) into decision-making processes while maintaining human dignity and autonomy. • Ethical management focuses on establishing shared responsibility among human users, AI developers, and those training the AI. This includes ensuring that human virtues are respected in AI design and application. • Boundaries: People may falsely attribute full moral agency to AI, despite its computational nature being fundamentally different from that of human moral reasoning.	• Ethical human-AI interaction focuses on ensuring that outcomes are shaped by continuous feedback loops between human and AI actions, thereby increasing overall beneficence. • Ethical management involves dynamically managing interactions to optimize outcomes for all stakeholders and aligning AI actions with human ethical goals. • Boundaries The unpredictable consequences of AI behavior in dynamic relationships can make it difficult to evaluate ethical outcomes clearly in advance.
Substantial Ontology (<i>Strong Sociomateriality</i>)	• Ethical human-AI interaction means that AI follows fixed rules, aligned with Kantian	• Ethical human-AI interaction assumes human virtues are stable and intrinsic, unaffected by	• Ethical human-AI interaction is defined by AI systems providing predefined outcomes

	deontology, where both AI and human actors adhere to stable ethical guidelines. Humans behave in norm-compliant ways while AI remains consistent with its programmed ethical principles. • Ethical management involves managing each entity separately, ensuring that AI systems are designed around universal ethical principles and that humans are trained for norm compliance. • Boundaries: This approach lacks flexibility, as AI cannot adapt to different contexts where multiple ethical principles might be necessary.	AI operations, while AI remains a neutral tool. The AI's connection to virtue ethics is based on the reflection of its developers' moral values and ethical decisions made during the design and training of the technology. • Ethical management focuses on cultivating human virtues independently of AI influence, while fostering ethical responsibility in developers during the design process. • Boundaries: A strict separation can lead to moral dilemmas, particularly when AI systems act autonomously, possibly diminishing human ethical decision-making or reducing engagement with virtue ethics.	based on their programming, which are designed for optimal performance and accuracy. These outcomes remain unaffected by individual interactions with human users, as they follow their own utilitarian goals to maximize task efficiency through AI usage. • Ethical management focuses on increasing the predictability and consistency of AI outputs, ensuring that these predefined outcomes maximize benefits for human users. • Boundaries: Predefined outcomes may fail to align with evolving social contexts, limiting the AI's ethical responsiveness and potentially leading to undesirable consequences in certain situations.

Table 4: Intersection and boundaries of ontologies and ethical theories in human-AI interaction.

As Table 4 demonstrates, both relational and substantial ontologies provide valuable yet incomplete frameworks for understanding ethical human-AI interaction. Traditional ethical theories, typically aligned with substantial ontology, ensure stability but may struggle with the dynamic and interactive nature of emerging AI technologies. Relational ontology offers flexibility and acknowledges the evolving interplay between human and AI moral agency, but it introduces complexities in maintaining ethical consistency. Thus, by combining normative ethical theories with sociomateriality discourse following two different ontological positions, this dissertation offers insights into how different ethical and ontological perspectives can shape our understanding of ethical human-AI interaction, providing a hybrid approach for the ethical management of human-AI interaction that incorporates the strengths of both ontologies. In doing so, this dissertation responds to the calls from Berente et al. (2019), Hagendorff (2020), Mirbabaie et al. (2022), and Siau and Wang (2020) for a general theory that addresses the complexities of ethics in the context of human-AI interactions. It, thereby, highlights the crucial role of the nested social subsystem following Leonardi's (2012) conceptualization of sociotechnical systems in forming ontological environments and ethical perspectives. It advances this conceptualization, as illustrated in Figure 1 (Section 2.2), by providing a more detailed description of the roles, mechanisms, and strategies that arise within the social subsystem and how this impacts the ethical management of human-AI interactions.

5.2 Practical Implications

This dissertation highlights the critical need to deepen our understanding of what constitutes ethical human-AI interaction (linked to RQ1) and to identify the management practices that shape such interactions within organizations (linked to RQ2). Two major practical implications arise from this research. First, it offers various approaches to address moral dilemmas that stem from conflicting ethical perspectives within organizations. Second, it underscores the significance of enhancing AI literacy among all employees to mitigate digital divides.

A key practical implication of this dissertation is the emphasis on how organizations can integrate ethical management practices into their AI deployment and utilization strategies. It addresses the challenges that arise when different organizational stakeholders—such as users, developers, and managers—hold divergent views on what constitutes ethical human-AI interaction. These differences in perspective can lead to conflicting strategies for managing these interactions ethically. Therefore, it is essential to incorporate diverse ethical viewpoints, as relying on the perspectives of only a single stakeholder group risks creating biased or incomplete solutions (e.g., Hammerschmidt, Hafner, et al., 2025; Heyder et al., 2023). Thus, organizations need to be aware of different perspectives that potentially clash (illustrating moral dilemmas) and should adopt a balanced approach to manage human-AI interaction ethically (e.g., Hammerschmidt, 2025; Hammerschmidt, Passlack, et al., submitted). Organizations must recognize the potential for these perspectives to clash, leading to moral dilemmas, and should adopt a balanced approach to align them effectively. Such an approach would involve combining top-down mechanisms, such as establishing clear ethical guidelines and standards for AI design and usage (aligned with deontological ethics), with bottom-up strategies that cultivate an ethical culture rooted in employees' personal moral values (reflecting virtue ethics) (e.g., Hammerschmidt, 2025; Hammerschmidt, Passlack, et al., submitted). By blending these strategies, organizations can foster ethical decision-making at all levels, creating a holistic framework where both organizational norms and individual moral reasoning inform the ethical deployment and use of AI. In this way, the dissertation contributes valuable insights into how organizations can integrate ethical considerations into their AI strategies, as called for by Alsheibani et al. (2020), Vial (2019), and Zhang et al. (2022).

To address these moral dilemmas by incorporating diverse ethical perspectives in AI strategies, the dissertation's papers outline several pathways for organizations. For example, **Papers I and VI** emphasize the need to establish AI ethics boards or committees at multiple levels within the organizational structure (e.g., involving developers at the unit level, leaders at the organizational level, and policymakers at the societal level). These boards would oversee AI deployment to ensure that systems align with both organizational values and ethical standards. **Papers V and VI** highlight the importance of fostering a workplace culture that encourages employees to voice ethical concerns about AI, enabling organizations to become aware of potential moral issues early on. Additionally, **Paper X** suggests that organizations implementing mechanisms to promote ethical human-AI interactions must

themselves behave ethically, for example, by communicating their AI strategies transparently and dealing fairly with potential impacts on employees' roles (e.g., changes to job contracts). Furthermore, **Paper VII** introduces scenario-based planning and regular ethical audits to anticipate and address moral dilemmas before and during AI deployment. These practices can help identify and resolve conflicts between ethical principles—such as fairness versus efficiency—by consulting a diverse range of stakeholders.

A second key practical implication of this dissertation pertains to the critical role of AI literacy in improving ethical human-AI interactions. By increasing employees' understanding of AI's capabilities and limitations, organizations can enable more informed decision-making, reduce the risks of AI misuse, and promote more ethical outcomes (Hammerschmidt, Stolz, et al., 2025). Equipping employees with this knowledge can help bridge AI-related inequalities, ensuring that all employees—not just those with technical expertise—are able to work effectively with AI systems. In doing so, organizations can reduce the digital divide caused by unequal access to AI knowledge and skills, ensuring that all employees, regardless of their backgrounds or positions, benefit equitably from AI technologies. This dissertation thus contributes to the literature on the digital divide by demonstrating how AI literacy can be leveraged to reduce inequalities, redefining traditional concepts of the digital divide, and offering concrete strategies for organizations to address these gaps (as emphasized by Carter et al., 2020; Choudrie et al., 2022; Schneider et al., 2023; Vassilakopoulou & Hustad, 2023).

To overcome skill-related gaps, the dissertation's papers outline several actionable strategies for organizations. For instance, **Paper V** suggests that organizations should encourage employees to critically reflect on the ethical implications of AI in their work. Integrating feedback loops into AI implementation processes can help users overcome potential misunderstandings, while leaders acting as role models can encourage ongoing critical evaluation, which in turn helps bridge gaps in AI literacy and prevents the digital divide from widening. **Paper III** identifies the differences between AI-literate and non-literate employees and recommends the creation of mentorship programs where AI-literate employees work alongside non-experts to improve AI literacy across all departments. This approach would ensure that the ethical implications of AI technologies are widely understood and addressed throughout the organization, fostering an inclusive environment where AI is accessible to all employees. Additionally, **Papers VIII** and **IX** propose that organizations should develop and implement comprehensive training programs to educate employees about how AI systems function, the biases inherent in these systems, and the ethical implications they carry.

By providing these practical pathways, this dissertation contributes valuable knowledge to organizations seeking to ethically navigate the complexities of AI deployment and management.

5.3 Methodological Implications

This dissertation illustrates how to enhance theorizing by combining mainly deductive insights from systematic literature reviews building upon well-established research (Paré et al., 2015; Templier & Paré, 2018) with explorative and inductive insights from empirical investigations (Sarker et al., 2018). This approach allows for the combination of previous ethical theories and sociomaterial conceptualizations to study new phenomena appearing in the IS field. In doing so, this dissertation does not solely consolidate insights from existing research (**Papers I, II, VI, VII, and VIII**) but also underpins the identified and explained relationships with empirical evidence (**Papers III, IV, V, IX, and X**). Thus, using this approach increases our understanding of ethical human-AI interaction and its management. Based on the four shifts in views of “Next-Generation Theorizing” by Burton-Jones et al. (2021, p. 303), the following are the possible steps that enhance theorizing in future mixed-method investigations, combining literature reviews with explorative qualitative studies.

First, to “appreciate the history of ideas in IS” (Burton-Jones et al., 2021, p. 306), this dissertation recommends extensive, systematic, and broad literature reviews to prioritize an appreciation for the historical context of ideas within the IS field. This approach helps to identify major concepts, such as determining virtue ethics, consequentialism, and deontology as the most important and well-established ethical theories (**Papers I and VI**), and emerging trends such as the dominance of deontological principles within AI ethics research (**Papers I and II**). The historical lens of this dissertation enables critical engagement with well-established theories while introducing new phenomena, such as contemporary challenges arising through increased AI deployment, including ethical principles that clash with moral values (**Paper VII**), or potential inequalities widening the divide between humans who can benefit from AI and those who cannot (**Paper VIII**).

Second, this dissertation shows that different sociomateriality frameworks and normative ethical theories can provide multiple perspectives on how we understand the ethical management of human-AI interaction, thus, using “[t]heory as a lens” (Burton-Jones et al., 2021, p. 306). Therefore, this dissertation encourages researchers to reflect on how their work represents, for instance, a normative ethical theory or a distinct ontological perspective, and how the chosen frameworks shape their analysis and interpretation (**Papers I and II**). Such reflexivity can lead to richer insights and highlight the importance of adapting methodologies to fit the complexities of the phenomena being studied.

Third, by combining systematic literature reviews with empirical investigations, this dissertation shows how to conceptualize theory as a “product of reflective engagement and action research.” (Burton-Jones et al., 2021, p. 306) By engaging with established theoretical constructs through literature reviews while simultaneously exploring real-world perceptions and capabilities through qualitative studies, this dissertation generates causal explanations that bridge theory and practice. This approach fosters a dynamic research process where insights from empirical data can inform and refine existing

sociomateriality theories (**Papers V and IX**), resulting in a more iterative and responsive approach to theorizing in the context of human-AI interactions.

Fourth, this dissertation is aware that sociomateriality is itself a phenomenon that shapes theorizing (Burton-Jones et al., 2021), particularly given its performative nature. The way this dissertation frames ethics and sociomateriality influences outcomes and interpretations of findings (**Papers VI and VII**). This dissertation, thus, aims to encourage researchers to conduct studies that not only capture the complexities of human-AI interactions but also consider the potential consequences of their theoretical framing.

In conclusion, this dissertation shows that when theorizing in the IS domain, Burton-Jones's (2021) four shifts in theorizing not only enhance methodological research but also allow for an in-depth examination of a phenomenon, such as the changing nature of human-AI interaction, based on the historical context of research on ethical theories and sociomateriality.

5.4 Limitations and Directions for Future Research

This dissertation has several limitations that also highlight potential directions for future research.

First, the focus of this thesis is primarily on organizations. However, much of the existing literature concerning perceptions and literacy, particularly in addressing RQ1, centers on children and students. Expanding the perspective to include societal dimensions could provide valuable insights. For instance, discussing the moral thresholds for acceptable behavior necessitates a more global perspective that integrates diverse cultural systems. Future research could examine how different cultural contexts influence perceptions of moral agency, particularly in relation to various regulatory initiatives aimed at addressing ethical issues surrounding AI such as reducing AI inequalities (see **Paper VIII**). This could involve comparative studies across regions to identify how local norms and values shape ethical expectations and behaviors in human-AI interactions. Additionally, incorporating a broader range of stakeholders, including community groups and nonprofit organizations, could enrich the discourse around ethical AI usage in various societal contexts.

Second, the samples utilized in this dissertation—both in the literature review and among interview participants—are restricted to specific domains. While the literature review draws from a wide array of databases, the ethical debates encountered in niche databases may differ significantly from those in broader contexts, potentially representing the virtue ethics of minority groups. Future research could benefit from employing alternative methodologies, such as snowball sampling, to gain insights from these underrepresented domains. Additionally, conducting longitudinal literature analyses would be valuable in illustrating trends over time, particularly since **Papers I and II** identify certain publication trends and citation network clusters. The participant pool for interviews primarily consisted of

individuals from organizational contexts, including employees, leaders, and developers. Future studies might explore entirely distinct demographic groups, such as parents and elderly individuals, to understand how different age groups perceive ethical requirements in private versus professional environments. The empirical findings in **Papers III, V, and X** indicate notable differences based on age and context, suggesting that these dimensions warrant further exploration.

Third, using the notions of sociomateriality to explore ethical interdependencies, particularly through the weak and strong forms, allows for an examination of different ontological positions regarding how ethics are perceived in human-AI interactions based on varying interpretations of human and AI moral agency. While these notions represent a valuable approach to explaining the interplay of several ethical perspectives that can arise in organizations, the results depend on them. Future work can use these notions to explain empirical interdependencies in relation to different ontological positions, for instance, in experimental studies on the interplay of different organizational stakeholders' ethical perceptions and behaviors with the ethical or nonethical outcomes of human-AI interactions. By exploring the interplay between material artifacts (AI technologies) and social structures (cultural norms, organizational policies, human behavior) through the sociomaterial notions, researchers can gain deeper empirical insights into how these ethical interdependencies shape ethical behavior based on one's ontological position.

Fourth, while this dissertation addresses meta-, applied, and descriptive ethics, it primarily focuses on normative ethics, particularly virtue ethics and deontology. This emphasis may limit the breadth of understanding regarding how ethical frameworks are perceived and enacted in practice. Future research should aim to develop methods for measuring normative ethical perspectives among different human stakeholders, facilitating empirical investigations into this topic. In particular, there is a significant gap in the literature regarding descriptive ethics, as illustrated in **Paper I**. By addressing this gap, future experimental studies can contribute to a more comprehensive understanding of how ethical principles are applied in real-world contexts and how they influence human-AI interactions.

In conclusion, acknowledging these limitations opens up numerous avenues for future research, emphasizing the need for a multifaceted approach that incorporates diverse perspectives, methodologies, and ethical frameworks to foster a deeper understanding of ethical human-AI interactions and their management.

6 Conclusion

This dissertation advances our understanding of ethical human-AI interactions and their management within organizations as sociotechnical systems. By employing the lens of sociomateriality and its dual ontological positions—relational and substantial ontology—the dissertation addresses two key research questions. First, it describes how human and AI capabilities, along with human perceptions of AI,

influence ethical human-AI interactions (RQ1) based on five articles (**Papers I to V**). Second, it explains how we can understand the ethical management of these interactions (RQ2) based on another five articles (**Papers IV to X**).

The need to investigate ethical human-AI interaction and its management arises from AI's increasing agency, which is transforming the nature of these interactions in organizations (Berente et al., 2019; Maedche et al., 2019; Rai et al., 2019). AI technologies no longer function merely as an extension of human intent, acting as moral patients (Allen et al., 2000; Hammerschmidt, Hafner, et al., 2025). Instead, they actively shape interactions, producing dynamic and sometimes unpredictable responses with moral implications for humans, assuming the role of moral agents (Allen et al., 2000; Hammerschmidt, Hafner, et al., 2025). This shift challenges traditional concepts of both human and material agency (Endacott & Leonardi, 2022; Hammerschmidt, Hafner, et al., 2025; Heyder et al., 2023). Thus, the main objective of this dissertation is to investigate the understanding and management of human-AI interaction from an ethical perspective, accounting for the changing dynamics between human and AI moral agencies. To address these goals, the methodological approach of this dissertation uses theoretical foundations derived from systematic literature reviews (**Papers I, II, VI, VII, and VIII**) and enhances them with empirical insights on how AI and human capabilities alter human perceptions of AI and moral agency (**Papers III, IV, V, IX, and X**), which reflects "Next-Generation Theorizing," as proposed by Burton-Jones et al. (2021, p. 303) and a nature of theory that focuses on explanation (Gregor, 2006).

By combining well-established normative ethical theories (such as deontology, virtue ethics, and consequentialism) with the notions of sociomateriality, this dissertation illustrates how diverse ethical and ontological perspectives inform our understanding of ethical human-AI interaction. Thereby, it extends the conceptualization of Heyder et al. (2023), which integrates the weak and strong notions of sociomateriality, following Jones (2014)—inseparability, relationality, performativity, materiality, and practices—into Leonardi's (2012) conceptualization of sociotechnical systems. By leveraging the weak and strong notions of sociomateriality, this dissertation provides a more nuanced analysis of the different ethical interdependencies between AI and human moral agency, demonstrating how human and AI capabilities reshape relationships, capabilities, and perceptions among various human stakeholders who engage with AI at different stages of the AI lifecycle.

Furthermore, it highlights the limitations of relying on these ethical theories in isolation, particularly in managing the complexities of human-AI interaction ethically. In doing so, it advocates for a more nuanced approach that accounts for potential conflicts between the different moral maxims of various human stakeholders from distinct moral communities and AI's ethical principles or norms and guidelines provided by the social subsystem. The dissertation reveals that the interplay of weak and strong sociomateriality, conceptualized as a sociomaterial cycle of interdependencies, is critical for managing human-AI interactions ethically in a manner that is sensitive to both relational and substantial

ontological perspectives. The findings suggest that ethical management must balance human oversight with AI's growing decision-making capabilities, ensuring that AI systems are developed with robust ethical guidelines while also empowering users to critically assess and intervene when necessary. Additionally, ethical management must clarify roles related to the distinction between human and AI moral agency to ensure that AI users do not misattribute moral responsibility to the technology. Furthermore, it must align their organizations' strategies with workforce capabilities and moral values to contribute to responsible and sustainable outcomes in diverse AI application contexts. This requires dialogue with diverse human stakeholders from the nested social subsystem.

Overall, this dissertation contributes to the growing body of research on the ethics of human-AI interaction (e.g., Brendel et al., 2021; Heyder et al., 2023; Mirbabaie et al., 2022), addressing the complex ethical challenges posed by AI's increasing adoption in organizations. As AI continues to develop, the boundaries between human intent and the technology's autonomy blur, prompting us to reconsider our fundamental understanding of agency and ethics. Ethical management is not solely about designing AI based on ethical principles; it is also about cultivating an ethical and AI-literate society that engages with these technologies in ways that elevate their benefits while also assessing their risks. Therefore, this dissertation lays the groundwork for future studies that delve deeper into the complex relationships between ontological perspectives, ethical theories, and sociomateriality, enhancing our understanding of ethics within human-AI interactions.

References

- Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk Research: Review and Recommendations. *Journal of Management*, 47(4), 823–837. <https://doi.org/10.1177/0149206320969787>
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251–261. <https://doi.org/10.1080/09528130050111428>
- Alsheibani, S. A., Cheung, Y., Messom, D. C., & Alhosni, M. (2020). Winning AI Strategy: Six-Steps to Create Value from Artificial Intelligence. *Americas Conference on Information Systems (AMCIS)*.
- Alt, R. (2020). Electronic Markets on sustainability. *Electronic Markets*, 30(4), 667–674. <https://doi.org/10.1007/s12525-020-00451-2>
- Anderson, M., & Anderson. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4).
- Aristoteles, Ostwald, M., & Aristoteles. (1962). *The Nicomachean ethics*. Prentice Hall.
- Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2020). Challenges of Explaining the Behavior of Black-Box AI Systems. *MIS Quarterly Executive*, 259–278. <https://doi.org/10.17705/2msqe.00037>
- Atkinson, R. F., Garner, R. T., & Rosen, B. (1969). Moral Philosophy: A Systematic Introduction to Normative Ethics and Meta-Ethics. *The Philosophical Quarterly*, 19(75), 181. <https://doi.org/10.2307/2217703>
- Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/MISQ/2021/15882>
- Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, 185(4), 725–740. <https://doi.org/10.1007/s10551-023-05339-7>
- Baptista, J., Stein, M.-K., Klein, S., Watson-Manheim, M. B., & Lee, J. (2020). Digital work and organisational transformation: Emergent Digital/Human work configurations in modern organisations. *The Journal of Strategic Information Systems*, 29(2), 101618. <https://doi.org/10.1016/j.jsis.2020.101618>
- Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, 28(3), 801–831. <https://doi.org/10.1086/345321>
- Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.
- Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, 30(2), 195–218. <https://doi.org/10.1007/s11023-020-09525-8>
- Bentham, J. (1988). *The principles of morals and legislation*. Batoche Books Kitchener.
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). MANAGING ARTIFICIAL INTELLIGENCE. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>

- Bilal, A., Wingreen, S., Sharma, R., & Jahanbin, P. (2021). Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics. *Proceedings of the Australasian Conference on Information Systems (ACIS)*.
- Bleher, H., & Braun, M. (2023). Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice. *Science and Engineering Ethics*, 29(3), 21. <https://doi.org/10.1007/s11948-023-00443-3>
- Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'Systematic' in Literature Reviews in IS. *Journal of Information Technology*, 30(2), 161–173. <https://doi.org/10.1057/jit.2014.26>
- Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines*, 27(4), 591–608. <https://doi.org/10.1007/s11023-017-9448-z>
- Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75(1), 653–675. <https://doi.org/10.1146/annurev-psych-030123-113559>
- Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly*, 25(1), 1. <https://doi.org/10.2307/3250956>
- Bowie, N. E. (1998). A Kantian Theory of Meaningful Work. *Journal of Business Ethics*, 17(9/10), Article 9/10. <https://doi.org/10.1556/084.2022.00108>
- Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. *Sustainability*, 13(4), 1974. <https://doi.org/10.3390/su13041974>
- Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26. <https://doi.org/10.1007/s10676-018-9448-6>
- Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. *Business Ethics Quarterly*, 33(1), 146–179. <https://doi.org/10.1017/beq.2021.42>
- Burton-Jones, A., Butler, B. S., Scott, S. V., & Xu, S. X. (2021). NEXT-GENERATION INFORMATION SYSTEMS THEORIZING: A CALL TO ACTION. *MIS Quarterly*, 45(1), 301–314.
- Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, 23(3), 387–399. <https://doi.org/10.1007/s11019-020-09948-1>
- Callon, M. (1999). Actor-Network Theory—The Market Test. *The Sociological Review*, 47(1_suppl), 181–195. <https://doi.org/10.1111/j.1467-954X.1999.tb03488.x>
- Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3290605.3300733>
- Carter, L., Liu, D., & Cantrell, C. (2020). Exploring the Intersection of the Digital Divide and Artificial Intelligence: A Hermeneutic Literature Review. *AIS Transactions on Human-Computer Interaction*, 12(4), 253–275. <https://doi.org/10.17705/1thci.00138>
- Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2), 74–78. <https://doi.org/10.1038/s42256-019-0020-9>
- Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The Sociomateriality of Information Systems. *MIS Quarterly*, 38(3), 809–830.

- Champagne, M., & Tonkens, R. (2023). A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot harm. *Science and Engineering Ethics*, 29(4), 27. <https://doi.org/10.1007/s11948-023-00449-x>
- Chedrawi, C., & Haddad, G. (2022). The Rise of Quasi-Humans in AI Fueled Organizations, an Ultimate socio-materiality approach to the Lens of Michel Serres. *Pacific Asia Journal of the Association for Information Systems*, 14(2).
- Chen, Y., Liu, H., & Chen, M. (2022). Achieving novelty and efficiency in business model design: Striking a balance between IT exploration and exploitation. *Information & Management*, 59(3), 103268. <https://doi.org/10.1016/j.im.2020.103268>
- Cheruvalath, R. (2023). Artificial Intelligent Systems and Ethical Agency. *Journal of Human Values*, 29(1), 33–47. <https://doi.org/10.1177/09716858221119546>
- Choudrie, J., Zamani, E., & Obuekwe, C. (2022). Bridging the Digital Divide in Ethnic Minority Older Adults: An Organisational Qualitative Study. *Information Systems Frontiers*, 24(4), 1355–1375. <https://doi.org/10.1007/s10796-021-10126-8>
- Churchill, L. R. (1999). Are we professionals? A critical look at the social role of bioethicists. *Daedalus*, 128, 253–274.
- Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Colmenarejo, A. B., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., & Kinder-Kurlanda, K. (2022). Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 107–118. <https://doi.org/10.1145/3514094.3534158>
- Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). <https://doi.org/10.1609/aaai.v31i1.11140>
- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*, 23(4), 803–814. <https://doi.org/10.1007/s10676-021-09616-9>
- Constantiou, I., Mayur, J., & Stelmaszak, M. (2023). Organizations as Digital Enactment Systems: A Theory of Replacement of Humans by Digital Technologies in Organizational Scanning, Interpretation, and Learning. *Journal of the Association for Information Systems*, 24(6), 1170–1798.
- Cook, T. (2023). Robust Artificial Moral Agents and Metanormativity. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 162–169. <https://doi.org/10.1145/3600211.3604703>
- Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022). Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 864–876. <https://doi.org/10.1145/3531146.3533150>
- Corbett, J. (2013). Designing and Using Carbon Management Systems to Promote Ecologically Responsible Behaviors. *Journal of the Association for Information Systems*, 14(7), 339–378. <https://doi.org/10.17705/1jais.00338>

Córdova, P. R., & Vicari, R. M. (2022). Practical Ethical Issues for Artificial Intelligence in Education. In *Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal, August 31–September 2, 2022, Revised Selected Papers* (Vol. 1720, pp. 437–453). Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-22918-3>

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed). Sage publ.

Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2019). Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. *Applied Artificial Intelligence*, 33(3), 267–293. <https://doi.org/10.1080/08839514.2018.1560124>

Darwall, S. (2005). Virtue Ethics. *Australasian Journal of Philosophy*, 83(4), 589–597. <https://doi.org/10.1080/00048400500339003>

Diederich, S., University of Göttingen, Germany, Brendel, A. B., TU Dresden, Germany, Morana, S., Saarland University, Germany, Kolbe, L., & University of Göttingen, Germany. (2022). On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *Journal of the Association for Information Systems*, 23(1), 96–138. <https://doi.org/10.17705/1jais.00724>

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>

Dyrkolbotn, S., Pedersen, T., & Slavkovik, M. (2018). On the Distinction between Implicit and Explicit Ethical Agency. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 74–80. <https://doi.org/10.1145/3278721.3278769>

Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1(1), 73–80. <https://doi.org/10.1007/s43681-020-00011-6>

Emdad, F. B., Ho, S. M., Ravuri, B., & Hussain, S. (2023). Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence. *Twenty-Ninth Americas Conference on Information Systems*.

Endacott, C. G., & Leonardi, P. M. (2022). Artificial Intelligence and Impression Management: Consequences of Autonomous Conversational Agents Communicating on One's Behalf. *Human Communication Research*, 48(3), 462–490. <https://doi.org/10.1093/hcr/hqac009>

Etikan, I. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1. <https://doi.org/10.11648/j.ajtas.20160501.11>

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, 21(4), 403–418. <https://doi.org/10.1007/s10892-017-9252-2>

Fabri, L., Häckel, B., Oberländer, A. M., Rieg, M., & Stohr, A. (2023). Disentangling Human-AI Hybrids: Conceptualizing the Interworking of Humans and AI-Enabled Systems. *Business & Information Systems Engineering*, 65(6), 623–641. <https://doi.org/10.1007/s12599-023-00810-1>

Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). <https://doi.org/10.1609/aaai.v31i1.10635>

Faulkner, P., & Runde, J. (2010). The Social, the Material, and the Ontology of Non-Material Technological Objects. *European Group for Organizational Studies (EGOS) Colloquium, Gothenburg*, 985, 4–8.

Figueras, C., Verhagen, H., & Pargman, T. C. (2022). Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations. *Scandinavian Journal of Management*, 34(2).

Flathmann, C., Schelble, B. G., Zhang, R., & McNeese, N. J. (2021). Modeling and Guiding the Creation of Ethical Human-AI Teams. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 469–479. <https://doi.org/10.1145/3461702.3462573>

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>

Forsyth, S., Dalton, B., Foster, E. H., Walsh, B., Smilack, J., & Yeh, T. (2021). Imagine a More Ethical AI: Using Stories to Develop Teens' Awareness and Understanding of Artificial Intelligence and its Societal Impacts. *2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, 1–2. <https://doi.org/10.1109/RESPECT51740.2021.9620549>

Fosso Wamba, S., Queiroz, M. M., Pappas, I. O., & Sullivan, Y. (2024). Artificial Intelligence Capability and Firm Performance: A Sustainable Development Perspective by the Mediating Role of Data-Driven Culture. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-023-10460-z>

Frankena, W. K. (1939). The naturalistic fallacy. *Mind*, 48(192), 464–477.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, 23(4), 791–801. <https://doi.org/10.1007/s10676-021-09615-w>

Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy*, 12(S6), 32–44. <https://doi.org/10.1111/1758-5899.12965>

Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic Book.

Gaus, G. F. (2001). What is Deontology? Part One: Orthodox View. *The Journal of Value Inquiry*, 35, Article 35. <https://doi.org/10.1145/3594536.3595156>

Gert, B. (2005). *Morality: Its Nature and Justification* (1st ed.). Oxford University Press New York. <https://doi.org/10.1093/0195176898.001.0001>

Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>

Gilligan, C. (2014). Moral injury and the ethic of care: Reframing the conversation about differences. *Journal of Social Philosophy*, 45(1).

Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI. *Journal of Business Ethics*, 178(4), 1027–1041. <https://doi.org/10.1007/s10551-022-05056-7>

Gogoshin, D. L. (2021). Robot Responsibility and Moral Community. *Frontiers in Robotics and AI*, 8, 768092. <https://doi.org/10.3389/frobt.2021.768092>

Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1), 100330. <https://doi.org/10.1016/j.ijme.2019.100330>

Gregor. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611. <https://doi.org/10.2307/25148742>

Grodziewicz, J. P., & Hohol, M. (2023). Waiting for a digital therapist: Three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry*, 14, 1190084. <https://doi.org/10.3389/fpsy.2023.1190084>

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, 29(2), 101614. <https://doi.org/10.1016/j.jsis.2020.101614>

Hafner, A., Hammerschmidt, T., & Stolz, K. (2024). Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review. *Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI)*. Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>

Hammerschmidt, T. (2025). Navigating the Nexus of Ethical Standards and Moral Values. *Ethics & Information Technology*, (EIT), 27(17), <https://doi.org/10.1007/s10676-025-09826-5>

Hammerschmidt, T., Hafner, A., Stolz, K., Passlack, N., Posegga, O., & Gerholz, K.-H. (2025). A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation. *Information Systems Frontiers (ISF)*, <https://doi.org/10.1007/s10796-02510596-0>.

Hammerschmidt, T., Passlack, N., & Posegga, O. (working paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment. [previously 2nd revision round at JSIS].

Hammerschmidt, T., Passlack, N., Stolz, K., & Posegga, O. (2024). AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitudes toward AI Assistants? *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*. Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii. <https://doi.org/10.24251/HICSS.2023.051>

Hammerschmidt, T., Stolz, K., & Posegga, O. (2025). Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI. *Behaviour of Information Technology*, (BIT), <https://doi.org/10.1080/0144929X.2025.2500451>

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation*, 4(1), 3–28. <https://doi.org/10.1017/S1867299X00002762>

Hauer, T. (2022). Incompleteness of moral choice and evolution towards fully autonomous AI. *Humanities and Social Sciences Communications*, 9(1), 38. <https://doi.org/10.1057/s41599-022-01060-4>

Henderson, J. C., & Venkatraman, H. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal*, 38(2.3), 472–484. <https://doi.org/10.1147/SJ.1999.5387096>

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3), 101772. <https://doi.org/10.1016/j.jsis.2023.101772>

- Heyder, T., & Posegga, O. (2021). Extending the Foundations of AI Literacy (Research-in-Progress). *Proceedings of the 46nd International Conference on Information Systems (ICIS)*. Proceedings of the 46nd International Conference on Information Systems (ICIS), Austin, Texas. https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/
- Höffe, O. (Ed.). (2013). *Einführung in die utilitaristische Ethik: Klassische und zeitgenössische Texte* (5., überarbeitete und erweiterte Auflage). A. Francke Verlag.
- Holford, W. D. (2022). An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control? *Journal of Business Ethics*, 176(1), 141–157. <https://doi.org/10.1007/s10551-020-04640-z>
- Hong, J.-W., Choi, S., & Williams, D. (2020). Sexist AI: An Experiment Integrating CASA and ELM. *International Journal of Human-Computer Interaction*, 36(20), 1928–1941. <https://doi.org/10.1080/10447318.2020.1801226>
- Horneber, D., & Laumer, S. (2023). Algorithmic Accountability. *Business & Information Systems Engineering*, 65(6), 723–730. <https://doi.org/10.1007/s12599-023-00817-8>
- Horváth, I. (2022). AI in interpreting: Ethical considerations. *Across Languages and Cultures*, 23, 1–13. <https://doi.org/10.1556/084.2022.00108>
- Hosmer, L. T. (1987). Ethical Analysis and Human Resource Management. *Human Resource Management*, 26, 313–330. <https://doi.org/10.1002/hrm.3930260302>
- Huang, Z., Shen, Y., Li, J., Fey, M., & Brecher, C. (2021). A Survey on AI-Driven Digital Twins in Industry 4.0: Smart Manufacturing and Advanced Robotics. *Sensors*, 21(19), 6340. <https://doi.org/10.3390/s21196340>
- Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in Human Behavior*, 132, 107253. <https://doi.org/10.1016/j.chb.2022.107253>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Jones, C., Parker, M., & Bos, R. T. (2005). *Business Ethics. A Critical Approach*. Routledge Taylor & Francis Group.
- Jones, M. (2014). A Matter of Life and Death: Exploring Conceptualizations of Sociomateriality in the Context of Critical Care. *MIS Quarterly*, 38(3), 895–925. <https://doi.org/10.25300/MISQ/2014/38.3.12>
- Kagan, S. (1992). The Structure of Normative Ethics. *Philosophical Perspectives*, 6, 223. <https://doi.org/10.2307/2214246>
- Kagan, S. (2018). *Normative Ethics*. Routledge. <https://doi.org/10.4324/9780429498657>
- Kant, I. (1991). *The metaphysics of morals*. Cambridge University Press.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Kegel, M. M., & Ghanem, L. (2024). You Did That on Purpose! An Investigation of the Knobe Effect in Human-Robot Interactions. *Proceedings of the 57th Hawaii International Conference on System Sciences*.

- Kieslich, K., Keller, B., & Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1), 205395172210929. <https://doi.org/10.1177/20539517221092956>
- Kitchenham, B., Budgen, D., & Pearl Brereton, O. (2011). Using mapping studies as the basis for further research – A participant-observer case study. *Information and Software Technology*, 53(6), 638–651. <https://doi.org/10.1016/j.infsof.2010.12.011>
- Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1), 7–15. <https://doi.org/10.1016/j.infsof.2008.09.009>
- Klumpp, M., & Zijm, H. (2019). Logistics Innovation and Social Sustainability: How to Prevent an Artificial Divide in Human–Computer Interaction. *Journal of Business Logistics*, 40(3), 265–278. <https://doi.org/10.1111/jbl.12198>
- Koltay, T. (2011). The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture & Society*, 33(2), 211–221. <https://doi.org/10.1177/0163443710393382>
- Koniakou, V. (2023). From the “rush to ethics” to the “race for governance” in Artificial Intelligence. *Information Systems Frontiers*, 25(1), 71–102. <https://doi.org/10.1007/s10796-022-10300-6>
- Krügel, S., Ostermaier, A., & Uhl, M. (2023). *The moral authority of ChatGPT* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2301.07098>
- Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.
- Latour, B., & Venn, C. (2002). Morality and Technology. *Theory, Culture & Society*, 19(5–6), 247–260. <https://doi.org/10.1177/026327602761899246>
- Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice*, 5(4), 379–393. <https://doi.org/10.1007/BF01059830>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>
- Lee, T. W., Mitchell, T. R., & Sablinski, C. J. (1999). Qualitative Research in Organizational and Vocational Psychology, 1979–1999. *Journal of Vocational Behavior*, 55(2), 161–187. <https://doi.org/10.1006/jvbe.1999.1707>
- Leonardi. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS Quarterly*, 35(1), 147. <https://doi.org/10.2307/23043493>
- Leonardi, P. M. (2012). Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How Are They Different? Do We Need Them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and Organizing* (pp. 24–48). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199664054.003.0002>
- Liu, M., Brynjolfsson, E., & Dowlatabadi, J. (2021). Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis. *Management Science*, 67(8), 4665–4685. <https://doi.org/10.1287/mnsc.2020.3721>
- Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, 122, 875–888. <https://doi.org/10.1016/j.jbusres.2019.10.006>

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3313831.3376727>

Mabaso, B. A. (2021). Computationally rational agents can be moral agents. *Ethics and Information Technology*, 23(2), 137–145. <https://doi.org/10.1007/s10676-020-09527-1>

Mackie, J. (1990). *Ethics: Inventing right and wrong*. Penguin.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376445>

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>

Maier, B. S., Jussupow, E., & Heinzl, A. (2019). GOOD, BAD, OR BOTH? MEASUREMENT OF PHYSICIAN'S AMBIVALENT ATTITUDES TOWARDS AI. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Marabelli, M., & Galliers, R. D. (2017). A reflection on information systems strategizing: The role of power and everyday practices. *Information Systems Journal*, 27(3), 347–366. <https://doi.org/10.1111/isj.12110>

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1), 71–87.

Markus, M. L., & Rowe, F. (2021). Guest Editorial: Theories of Digital Transformation: A Progress Report. *Journal of the Association for Information Systems*, 22(2), 273–280. <https://doi.org/10.17705/1jais.00661>

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, 160(4), 835–850. <https://doi.org/10.1007/s10551-018-3921-3>

Mayer, A.-S., Strich, F., University of Passau (Germany), Fiedler, M., & University of Passau (Germany). (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, 239–257. <https://doi.org/10.17705/2msqe.00036>

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt.

McDonald, N., & Pan, S. (2020). Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–19. <https://doi.org/10.1145/3415218>

McGrath, S. K., & Whitty, S. J. (2018). Accountability and responsibility defined. *International Journal of Managing Projects in Business*, 11(3), 687–707. <https://doi.org/10.1108/IJMPB-06-2017-0058>

McHuhg, M. L. (2012). Interrater Reliability: The Kappa Statistic. *Lessons in Biostatistics*, 22(3), Article 3. <https://doi.org/10.1145/3600211.3604680>

McLean, K. (2013). Literacy and Technology in the Early Years of Education: Looking to the Familiar to Inform Educator Practice. *Australasian Journal of Early Childhood*, 38(4), 30–41. <https://doi.org/10.1177/183693911303800405>

- Meijer, A., Wiarda, M., Doorn, N., & Van De Kaa, G. (2023). Towards responsible standardisation: Investigating the importance of responsible innovation for standards development. *Technology Analysis & Strategic Management*, 1–15. <https://doi.org/10.1080/09537325.2023.2225108>
- Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3), 103434. <https://doi.org/10.1016/j.im.2021.103434>
- Mill, J. S., Bentham, J., & Austin, J. (1989). *Utilitarianism* (M. Warnock, Ed.; 19. impr). Fontana Press.
- Millar, J. (2016). An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars. *Applied Artificial Intelligence*, 30(8), 787–809. <https://doi.org/10.1080/08839514.2016.1229919>
- Mirbabaie, M., Brendel, A. B., Faculty of Business and Economics, Technische Universität Dresden, Hofeditz, L., & Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1), 726–753. <https://doi.org/10.17705/1CAIS.05034>
- Mom, T. J. M., Chang, Y.-Y., Cholakova, M., & Jansen, J. J. P. (2019). A Multilevel Integrated Framework of Firm HR Practices, Individual Ambidexterity, and Organizational Ambidexterity. *Journal of Management*, 45(7), 3009–3034. <https://doi.org/10.1177/0149206318776775>
- Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE INTELLIGENT SYSTEMS*.
- Morton, J., Amrollahi, A., & Wilson, A. D. (2022). Digital strategizing: An assessing review, definition, and research agenda. *The Journal of Strategic Information Systems*, 31(2), 101720. <https://doi.org/10.1016/j.jsis.2022.101720>
- Mutch, A. (2010). Technology, Organization, and Structure—A Morphogenetic Approach. *Organization Science*, 21(2), 507–520. <https://doi.org/10.1287/orsc.1090.0441>
- Mutch, A. (2013). Sociomateriality—Taking the wrong turning? *Information and Organization*, 23(1), 28–40. <https://doi.org/10.1016/j.infoandorg.2013.02.001>
- Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, 17(1), 2–26. <https://doi.org/10.1016/j.infoandorg.2006.11.001>
- Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, 71(1), 137–161. <https://doi.org/10.1007/s11423-023-10203-6>
- Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3), 336–359. <https://doi.org/10.1057/ejis.2012.26>
- Niederman, F., & Baker, E. W. (2023). Ethics and AI Issues: Old Container with New Wine? *Information Systems Frontiers*, 25(1), 9–28. <https://doi.org/10.1007/s10796-022-10305-1>
- Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-023-01635-y>
- O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 64(5), 904–938. <https://doi.org/10.1177/0018720820960865>

O'Reilly, C. A., & Tushman, M. L. (2004). The ambidextrous organization. *Harvard Business Review*, 82(4), 74–83.

O'Reilly, C. A., & Tushman, M. L. (2011). Organizational Ambidexterity in Action: How Managers Explore and Exploit. *California Management Review*, 53(4), 5–22. <https://doi.org/10.1525/cmr.2011.53.4.5>

Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. *Organization Studies*, 28(9), 1435–1448. <https://doi.org/10.1177/0170840607081138>

Orlikowski, W. J. (2010). The sociomateriality of organisational life: Considering technology in management research. *Cambridge Journal of Economics*, 34(1), 125–141. <https://doi.org/10.1093/cje/bep058>

Orlikowski, W. J., & Scott, S. V. (2008). 10 Sociomateriality: Challenging the Separation of Technology, Work and Organization. *Academy of Management Annals*, 2(1), 433–474. <https://doi.org/10.5465/19416520802211644>

Othman, U., & Yang, E. (2023). Human–Robot Collaborations in Smart Manufacturing Environments: Review and Outlook. *Sensors*, 23(12), 5663. <https://doi.org/10.3390/s23125663>

Pansoni, S., Tiribelli, S., Paolanti, M., Di Stefano, F., Frontoni, E., Malinverni, E. S., & Giovanola, B. (2023). ARTIFICIAL INTELLIGENCE AND CULTURAL HERITAGE: DESIGN AND ASSESSMENT OF AN ETHICAL FRAMEWORK. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2–2023, 1149–1155. <https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-1149-2023>

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>

Peppard, J., & Ward, J. (2004). Beyond strategic information systems: Towards an IS capability. *The Journal of Strategic Information Systems*, 13(2), 167–194. <https://doi.org/10.1016/j.jsis.2004.02.002>

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1), 34–47. <https://doi.org/10.1109/TTS.2020.2974991>

Petter, Straub, & Rai. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, 31(4), 623. <https://doi.org/10.2307/25148814>

Pickering, A. (1993). The Mangle of Practice: Agency and Emergence in the Sociology of Science. *American Journal of Sociology*, 99(3), 559–589. <https://doi.org/10.1086/230316>

Pinski, M., & Benlian, A. (2023). AI Literacy—Towards Measuring Human Competency in Artificial Intelligence. *Hawaii International Conference on System Sciences*.

Pinski, M., Hofmann, T., & Benlian, A. (2024). AI Literacy for the top management: An upper echelons perspective on corporate AI orientation and implementation ability. *Electronic Markets*, 34(1), 24. <https://doi.org/10.1007/s12525-024-00707-1>

Polyviou, A., & Zamani, E. D. (2022). TOWARDS EUROPE'S AI STRATEGY: DESIRES & REALITIES. *Thirtieth European Conference on Information Systems*.

Puri, A. (2020). Moral Imitation: Can Algorithm Really Be Ethical? *Rutgers Law Record*, 48(1), 47–57.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, 43(1), iii–ix.

Raisch, S., & Krakowski, S. (2021). ARTIFICIAL INTELLIGENCE AND MANAGEMENT: THE AUTOMATION-AUGMENTATION PARADOX. *Academy of Management Review*, 46(1), 192–210. <https://doi.org/10.5465/amr.2018.0072>

Rawls, J. (1971). *A Theory of Justice*. Cambridge (Mass.).

Riggins, F., & Dewan, S. (2005). The Digital Divide: Current and Future Research Directions. *Journal of the Association for Information Systems*, 6(12), 298–337. <https://doi.org/10.17705/1jais.00074>

Roberts, H., Cows, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2023). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, 39(2), 79–97. <https://doi.org/10.1080/01972243.2022.2124565>

Roberts, J. S., & Montoya, L. N. (2022). Contextualizing Artificially Intelligent Morality: A Meta-Ethnography of Top-Down, Bottom-Up, and Hybrid Models for Theoretical and Applied Ethics in Artificial Intelligence. *Future of Information and Communication Conference*, *arXiv:2204.07612*. <http://arxiv.org/abs/2204.07612>

Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, 33(1), 100925. <https://doi.org/10.1016/j.hrmr.2022.100925>

Saariluoma, P., & Leikas, J. (2020). Designing Ethical AI in the Shadow of Hume’s Guillotine. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), *Intelligent Human Systems Integration 2020* (Vol. 1131, pp. 594–599). Springer International Publishing. https://doi.org/10.1007/978-3-030-39512-4_92

Sarker, S., Xiao, X., Copenhagen Business School, Beaulieu, T., Utah State University, Lee, A. S., & Virginia Commonwealth University. (2018). Learning from First-Generation Qualitative Approaches in the IS Discipline: An Evolutionary View and Some Implications for Authors and Evaluators (PART 1/2). *Journal of the Association for Information Systems*, 19, 752–774. <https://doi.org/10.17705/1jais.00508>

Schlagwein, D., & Willcocks, L. (2023). ‘ChatGPT et al.’: The ethics of using (generative) artificial intelligence in research and science. *Journal of Information Technology*, 38(3), 232–238. <https://doi.org/10.1177/02683962231200411>

Schneider, J., Meske, C., & Bikic, A. (2023). How individuals can shape AI through data—An AI literacy and morality perspective. *Thirty-First European Conference on Information Systems*.

Scott, S. V., & Orlikowski, W. J. (2013). Sociomateriality — taking the wrong turning? A response to Mutch. *Information and Organization*, 23(2), 77–80. <https://doi.org/10.1016/j.infoandorg.2013.02.003>

Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7(1), 119. <https://doi.org/10.1057/s41599-020-00614-8>

Shahid, I., Anwar Ul Haq, M., Di Virgilio, F., & Ahmed, M. A. (2023). Ethical leadership and individual ambidexterity: The mediating role of equity sensitivity and the moderating role of openness to experience. *Management Research Review*, 46(12), 1727–1747. <https://doi.org/10.1108/MRR-06-2022-0418>

- Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22(5), 648–663. <https://doi.org/10.1080/1369118X.2019.1568515>
- Shank, D. B., & Gott, A. (2020). Exposed by AIs! People Personally Witness Artificial Intelligence Exposing Personal Information and Exposing People to Undesirable Content. *International Journal of Human–Computer Interaction*, 36(17), 1636–1645. <https://doi.org/10.1080/10447318.2020.1768674>
- Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31. <https://doi.org/10.1145/3419764>
- Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, 31(2), 74–87. <https://doi.org/10.4018/JDM.2020040105>
- Simpson, P. (1992). Contemporary Virtue Ethics and Aristotle. *The Review of Metaphysics*, 45(3), 503–524.
- Singer, P. (1968). *Applied Ethics*. Oxford University Press.
- Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics*, 9(3–4), 331–352. <https://doi.org/10.1007/s10892-005-3508-y>
- Sirén, C. A., Kohtamäki, M., & Kuckertz, A. (2012). Exploration and exploitation strategies, profit performance, and the mediating role of strategic learning: Escaping the exploitation trap. *Strategic Entrepreneurship Journal*, 6(1), 18–41. <https://doi.org/10.1002/sej.1126>
- Søvik, A. O. (2022). How a non-conscious robot could be an agent with capacity for morally responsible behaviour. *AI and Ethics*, 2(4), 789–800. <https://doi.org/10.1007/s43681-022-00140-0>
- Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. <https://doi.org/10.1016/j.ijinfomgt.2023.102700>
- Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, 45, 105661. <https://doi.org/10.1016/j.clsr.2022.105661>
- Straub, D., & Gefen, D. (2004). Validation Guidelines for IS Positivist Research. *Communications of the Association for Information Systems*, 13. <https://doi.org/10.17705/1CAIS.01324>
- Strich, F., University of Bayreuth, Germany, Mayer, A.-S., University of Passau, Germany, Fiedler, M., & University of Passau, Germany. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees’ Professional Role Identity. *Journal of the Association for Information Systems*, 22(2), 304–324. <https://doi.org/10.17705/1jais.00663>
- Suárez-Gonzalo, S., Mas-Manchón, L., & Guerrero-Solé, F. (2019). Tay is you. The attribution of responsibility in the algorithmic culture. *Observatorio (OBS*)*, 13(2). <https://doi.org/10.15847/obsOBS13220191432>
- Susser, D. (2019). Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 403–408. <https://doi.org/10.1145/3306618.3314286>
- Swanepoel, D. (2021). The possibility of deliberate norm-adherence in AI. *Ethics and Information Technology*, 23(2), 157–163. <https://doi.org/10.1007/s10676-020-09535-1>

Tajalli, P. (2021). AI Ethics and the Banality of Evil. *Ethics and Information Technology*, 23, 447–454. <https://doi.org/10.1007/s10676-021-09587-x>

Templier, M., & Paré, G. (2018). Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, 27(5), 503–550. <https://doi.org/10.1080/0960085X.2017.1398880>

Terzis, P. (2020). Onward for the freedom of others: Marching beyond the AI ethics. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3351095.3373152>

Tóth, Z., Caruana, R., Gruber, T., & Loebbecke, C. (2022). The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability. *Journal of Business Ethics*, 178(4), 895–916. <https://doi.org/10.1007/s10551-022-05050-z>

Townsend, B., Paterson, C., Arvind, T. T., Nemirovsky, G., Calinescu, R., Cavalcanti, A., Habli, I., & Thomas, A. (2022). From Pluralistic Normative Principles to Autonomous-Agent Rules. *Minds and Machines*, 32(4), 683–715. <https://doi.org/10.1007/s11023-022-09614-w>

Unver, M. B. (2023). Reebuilding ‘ethics’ to Govern AI: How to re-set the boundaries for the legal sector? *International Conference on Artificial Intelligence and Law (ICAIL)*. <https://doi.org/10.1145/3594536.3595156>

Van Deursen, A., & Van Dijk, J. (2011). Internet skills and the digital divide. *New Media & Society*, 13(6), 893–911. <https://doi.org/10.1177/1461444810386774>

Vanhée, L., & Borit, M. (2022). Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence. *Journal of Artificial Intelligence Research*, 73, 619–631. <https://doi.org/10.1613/jair.1.13135>

Vassilakopoulou, P., & Hustad, E. (2023). Bridging Digital Divides: A Literature Review and Research Agenda for Information Systems Research. *Information Systems Frontiers*, 25(3), 955–969. <https://doi.org/10.1007/s10796-020-10096-3>

VHB. (2015). *VHB-JOURQUAL3 Ranking*. Verband der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft e.V. (VHB). <https://vhbonline.org/service/vhb-jourqual/vhb-jourqual-3/gesamtlste>

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>

Wang, Y., Xiong, M., & Olya, H. (2020). Toward an Understanding of Responsible Artificial Intelligence Practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Proceedings of the 53rd Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2020.610>

Weaver, G. R., & Trevino, L. K. (1994). Normative And Empirical Business Ethics: Separation, Marriage Of Convenience, Or Marriage Of Necessity? *Business Ethics Quarterly*, 4(2), 129–143. <https://doi.org/10.2307/3857485>

Wehrens, R., Stevens, M., Kostenzer, J., Weggelaar, A. M., & De Bont, A. (2023). Ethics as Discursive Work: The Role of Ethical Framing in the Promissory Future of Data-driven Healthcare Technologies. *Science, Technology, & Human Values*, 48(3), 606–634. <https://doi.org/10.1177/01622439211053661>

Wei, K.-K., Teo, H.-H., Chan, H. C., & Tan, B. C. Y. (2011). Conceptualizing and Testing a Social Cognitive Model of the Digital Divide. *Information Systems Research*, 22(1), 170–187. <https://doi.org/10.1287/isre.1090.0273>

White, J., Brandt, D., Söffner, J., & Stapleton, L. (2023). Artificial thinking and doomsday projections: A discourse on trust, ethics and safety. *AI & SOCIETY*, 38(6), 2119–2124. <https://doi.org/10.1007/s00146-023-01810-1>

Willcocks, L. (2020). Robo-Apocalypse cancelled? Reframing the automation and future of work debate. *Journal of Information Technology*, 35(4), 286–302. <https://doi.org/10.1177/0268396220925830>

Wilson, R. A., & Land, M. K. (2021). Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review*, 52(3), 1029–1194.

Winter, S., Berente, N., Howison, J., & Butler, B. (2014). Beyond the organizational ‘container’: Conceptualizing 21st century sociotechnical work. *Information and Organization*, 24(4), 250–269. <https://doi.org/10.1016/j.infoandorg.2014.10.003>

Yin, R. K. (1994). Discovering the Future of the Case Study. Method in Evaluation Research. *Evaluation Practice*, 15(3), 283–290. <https://doi.org/10.1177/109821409401500309>

Yu, B., Vahidov, R., & Kersten, G. E. (2021). Acceptance of technological agency: Beyond the perception of utilitarian value. *Information & Management*, 58(7), 103503. <https://doi.org/10.1016/j.im.2021.103503>

Zarnoufi, R., Boutbi, M., & Abik, M. (2020). AI to Prevent Cyber-Violence: Harmful Behaviour Detection in Social Media. *International Journal of High Performance Systems Architecture*, 9(4), 182–191.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *Journal of Artificial Intelligence Research*, 71. <https://doi.org/10.1613/jair.1.12895>

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101, 104327. <https://doi.org/10.1016/j.jesp.2022.104327>

Zheng, F., Zhao, C., Usman, M., & Poulouva, P. (2024). From Bias to Brilliance: The Impact of Artificial Intelligence Usage on Recruitment Biases in China. *IEEE Transactions on Engineering Management*, 71, 14155–14167. <https://doi.org/10.1109/TEM.2024.3442618>

Appendices

Appendix I: Paper

Table I-1 in Appendix I provides all ten articles of this dissertation (**Paper I to X**) in the following order:

Paper I	Hammerschmidt, T.; Hafner, A., Stolz, K.; Passlack, N., Posegga, O. & Gerholz, K-H. (2025). <i>A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation</i> . Information Systems Frontiers (ISF), https://doi.org/10.1007/s10796-025-10596-0 .
Paper II	Hafner, A., Hammerschmidt, T. & Stolz, K. (2024). <i>Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review</i> . Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany, https://aisel.aisnet.org/wi2024/46 .
Paper III	Hammerschmidt, T., Passlack, N., Stolz, K. & Posegga, O. (2024). <i>AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants?</i> Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, https://hdl.handle.net/10125/106426 .
Paper IV	Hammerschmidt, T., Stolz, K. & Posegga, O. (2024). <i>How Leaders' Ambidexterity and Literacy Impact AI Capabilities of Organizations</i> . Proceedings of 32nd the European Conference on Information Systems (ECIS), Paphos, Cyprus, https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14/ .
Paper V	Heyder, T. & Posegga (2021). <i>Extending the Foundations of AI Literacy</i> (Short Paper, Best Paper Nominee). Proceedings of the 46nd International Conference on Information Systems (ICIS), Austin, Texas, https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/ .
Paper VI	Heyder, T; Passlack, N. & Posegga, O. (2023) <i>Ethical Management of Human-AI Interaction: A Theory Development Review</i> . The Journal of Strategic Information Systems (JSIS), https://doi.org/10.1016/j.jsis.2023.101772 .
Paper VII	Hammerschmidt, T. (2025) <i>Navigating the Nexus of Ethical Standards and Moral Values</i> . Ethics and Information Technology (EIT), 27(17), https://doi.org/10.1007/s10676-025-09826-5 .
Paper VIII	Hammerschmidt, T.; Stolz, K. & Posegga, O. (2025). <i>Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI</i> . Behaviour and Information Technology, (BIT), https://doi.org/10.1080/0144929X.2025.2500451 .
Paper IX	Hammerschmidt, T.; Passlack; N. & Posegga, O. (working paper, unpublished). <i>How Organizations Ethically Respond to Decision Dilemmas in AI Deployment</i> . [previously 2 nd revision round at JSIS]
Paper X	Stolz, K., Hammerschmidt T. & Posegga, O. (2024). <i>How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making</i> (Best Paper Nominee). Proceedings of 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, https://hdl.handle.net/10125/106970 .

Table I-1: Overview of this dissertation's articles.

Paper I

Hammerschmidt, T.; Hafner, A.; Stolz, K.; Passlack, N.; Posegga, O. & Gerholz, K-H. (2025). *A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation*. Information Systems Frontiers (ISF), <https://doi.org/10.1007/s10796-025-10596-0>.

Information Systems Frontiers (ISF) Manuscript Submission titled

**A Review of How Different Views on Ethics Shape
Perceptions of Morality and Responsibility within AI Transformation**

for the Special Issue “Artificial Intelligence and Machine Learning: New Frontier of
Research, Practice and Method for Transformative Changes in IS”

Teresa Hammerschmidt (née Heyder; first and corresponding author)

Affiliation: University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany, +49 176
31303413, teresa.heyder@uni-bamberg.de, ORCID: <https://orcid.org/0000-0003-2164-3825>

Alina Hafner (second author)

Affiliation: TUM School of Computation, Information and Technology, Technical University of
Munich, Arcisstraße 21, 9, 80333 Munich, Germany, alina.hafner@tum.de, ORCID:
<https://orcid.org/0009-0003-6236-3018>

Katharina Stolz (third author)

Affiliation: University of Stuttgart, Keplerstraße 17, 70174 Stuttgart, Germany,
katharina.stolz@bwi.uni-stuttgart.de, ORCID: <https://orcid.org/0000-0002-4315-7361>

Nina Passlack (fourth author)

Affiliation: University of Bamberg, Kärntenstraße 7, 96052 Bamberg, Germany,
nina.passlack@uni-bamberg.de, ORCID: <https://orcid.org/0000-0002-2784-7645>

Oliver Posegga (fifth author)

Affiliation: University of Bamberg, An der Weberei 5, Bamberg, Germany, oliver.posegga@uni-bamberg.de, ORCID: <https://orcid.org/0000-0003-4552-3027>

Karl-Heinz Gerholz (sixth author)

Affiliation: University of Bamberg, Kärntenstraße 7, 96052 Bamberg, Germany, karl-heinz.gerholz@uni-bamberg.de, ORCID: <https://orcid.org/0000-0003-2118-4091>

A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation

Research Paper – Accepted Version

Abstract

This paper examines the evolving dynamics of human-AI interaction, emphasizing the ethical challenges and responsibility gaps that emerge as AI technologies become more autonomous and integrated into society and business. We analyze, utilizing a systematic literature review, how various ethical views influence our understanding of morality and responsibility in human-AI collaborations. Deontological ethics emerge as a dominant theme, with much of the literature centered on ethical principles shaped by powerful nations. The study highlights the need to integrate diverse ethical perspectives into AI research to address contradictions in ethical frameworks across various cultural contexts. While respecting cultural differences, achieving a common ground among these frameworks requires increased dialogue among AI researchers and practitioners. Our findings further underscore the importance of future research in developing a more cohesive understanding of how AI transformation challenges previous assumptions about AI's role in moral agency and responsibility.

Keywords: *Ethical Theories, Responsible AI, Moral Philosophy, Systematic Review*

Introduction

The advent of recent generations of artificial intelligence (AI) technologies, characterized by self-learning, human-like, and (semi-)autonomous capabilities that emulate cognitive functions traditionally associated with the human mind, has profoundly transformed the dynamics of human-technology interaction (Dwivedi et al., 2023; Maedche et al., 2019; Rai et al., 2019). Today's AI exhibits greater interactivity and complexity levels than earlier technologies, thereby shifting agency from humans to these AI technologies (Baird & Maruping, 2021; Berente et al., 2019; Mirbabaie et al., 2022). As AI becomes increasingly integrated into various aspects of contemporary society and business, transforming how we generate knowledge, prepare decisions, and carry out tasks and actions, new responsibility gaps and ethical tensions emerge (Constantinescu et al., 2021; Dwivedi et al., 2021; Giermindl et al., 2022). These arise from the blurred interdependencies between data, AI, and the diverse roles humans play (e.g., developers, users, managers, regulators) and their accountabilities for unintended consequences (Giermindl et al., 2022; Horneber & Laumer, 2023).

This prompts critical questions about how we can understand morality and ethics in the context of human-AI interaction and address the arising responsibility gaps (Martinho et al., 2021). Who decides, for instance, what is right or wrong when decisions are made within human-AI teamwork? Can and should AI agents exhibit moral behavior akin to their human counterparts? These questions touch on moral uncertainties that could have significant and unforeseen implications for individuals, organizations, and societies (Dennehy et al., 2023; Dwivedi et al., 2023; Pappas et al., 2023). Thus, promoting critical inquiries into delineating responsibilities and attribution of agency represents an emerging challenge in AI transformation. There is an unfolding academic discourse on responsible and human-centered AI, which seeks to propose practices for designing, using, managing, and governing AI in a manner that accounts for the values and perceptions of those affected by AI transformation, ensuring moral and responsible collaboration between humans and AI (Pappas et al., 2023; Schmager et al., 2023; Vassilakopoulou et al., 2022).

One's understanding of morality can be explained through diverse ethical views (Vanhée & Borit, 2022). The ethical view of people who assess the rightness or wrongness of human-AI interactions based on whether the outcomes yield the greatest benefits for most people, for instance, can be described through a utilitarian perspective (Allen et al., 2000). Conversely, those who evaluate whether an AI's action meets the "become a universal law" criterion might apply Kant's categorical imperative, which can be explained with deontological ethics (Nowak, 2018, p. 189). While most individuals may not consciously adhere to a single ethical view, which is not necessarily mutually exclusive, different ethical perspectives can illuminate how ethical considerations are applied in human-AI interactions and how this impacts our understanding of morality and responsibilities. Clarifying these perspectives is essential to reduce uncertainties regarding responsibilities and accountabilities within the human-AI interaction (Figueras et al., 2022).

In this context, researchers are discussing normative approaches to moral and ethical agency by examining how both human and material entities can involve ethics and morality as part of their agency and how these entities construct their realities of ethical reflection and moral values (e.g., Behdadi & Munthe, 2020; Champagne & Tonkens, 2023; Floridi & Sanders, 2004; Niederman & Baker, 2023). Moral agency pertains to the capacity to take responsibility for actions and decisions based on the often invisible background of one's belief system, which is shaped by collective moral perceptions and societal values, establishing a shared understanding of what is "good" and "bad." (Mabaso, 2021) In contrast, ethical agency entails a more profound reflective process upon diverse moral maxims involving ethical assumptions as assumed beliefs to discern what is "right" from what is "wrong" (Mabaso, 2021), which can be interpreted through various ethical views, often referred to as "ethical theories" or "ethical families" (e.g., Bankins & Formosa, 2023; Bilal et al., 2021; Floridi & Sanders, 2004; Hindocha & Badea, 2022).

Capturing the concepts of agency and responsibility within the context of AI transformation is challenging, particularly as artificial agents extend traditional boundaries of moral involvement, highlighting the need for further research on understanding morality within human-AI teams (Behdadi & Munthe, 2020; Floridi & Sanders, 2004; Niederman & Baker, 2023). Therefore, the motivation for this work is, to analyze which ethical views are employed when examining human-AI interaction from an ethical standpoint and how this impacts our understanding of morality and responsibility. We aim to shed light on areas of moral uncertainty that arise as part of AI transformation, which are changing agency perceptions, particularly how humans interact with technologies

and assign responsibilities based on their inherent ontological view of morality within contexts such as AI design, deployment, curricula, regulation, and research (Bogosian, 2017; Rismani & Moon, 2023). Our research questions (RQs) are:

RQ1: *What ethical perspectives are addressed in AI ethics research?*

RQ2: *What challenges for establishing ethical frameworks are emerging within AI ethics research?*

RQ3: *In what ways might AI transformation challenge traditional notions of moral agency and responsibility?*

This work makes contributions by expanding the AI ethics discourse, combining technical discourse on the unique and novel capabilities of AI technologies with moral philosophy on ethics, agency, and responsibility, which is currently dispersed across several research disciplines (Berente et al., 2019; Fosso Wamba & Queiroz, 2023; Islam & Greenwood, 2021). We illustrate how the distinct self-learning, partly autonomous, and human-like capabilities of AI technologies, compared to previous forms of technologies, necessitate their own ethical stance, addressing similar ethical questions across a wide range of AI applications, from generative AI tools to robots. Thereby, we follow the arguments of Floridi and Sanders (2004), noting that artificial agents extend the traditional boundaries of moral involvement, as they can be considered both moral patients (entities that can be affected by actions) and moral agents (entities capable of performing actions that have moral implications) and as their moral implications can be quantified numerically, which challenges traditional concepts of morality, responsibility, and ethics in human-technology interactions. We also emphasize the diversity of normative ethical dialogue within more ethics-related research fields, despite the predominance of deontological ethics in Information System (IS) research, which often focuses on ethical principles guiding AI design and governance, such as the European Unions (EU) AI Act's attempt to clarify responsibilities. Finally, we analyze the publication dynamics surrounding researchers' interpretation of ethics and morality in the context of human-AI interaction, which has not yet been explored in detail.

Section 2 outlines the theoretical background. We describe the method of our literature review in Section 3: we use a descriptive review because it aims to structure studies in order to identify and interpret patterns and trends (Paré et al., 2015). Our results are summarized in Section 4. Section 5 discusses theoretical and practical implications, directions for future work, and limitations. This paper ends with a conclusion in Section 6.

Theoretical Background

This section offers an overview of different ethical views of normative ethics (Section 2.1) and examines the relationship between ethics, morality, and agency in the context of AI (Section 2.2).

2.1 Understanding Morality: Framing through Different Ethical Theories

Research on AI ethics involves ethical considerations that represent either indirectly or directly different views on ethics, also sometimes referred to as “ethical families” or “ethical theories” (e.g., Bankins & Formosa, 2023; Bilal et al., 2021; Floridi & Sanders, 2004; Hindocha & Badea, 2022). Ethics, which is understood as the science of morality, can be defined as “the capacity to think critically about moral values and direct our actions in terms of such values.” (Churchill, 1999, p. 259). While ethical theories often engage with abstract views on what is good or bad related to societal moral values and cultural norms, they also offer frameworks for understanding how these moral values translate into direct actions and evaluating the consequences of actions in an ethical manner (Bryson,

2018; Lim & Kwon, 2021). Within the context of information technologies, such as AI systems, ethical theories reflect upon the benefits and harms that can arise during human-AI interaction or as outcomes of this interaction.

Morality encompasses the prevailing notions of “right” and “wrong” within a particular context, society, and time, defined by a set of established norms and rules that are both explicit and implicit. These norms are based on social processes that vary over time and reflect the belief systems of various social entities, such as religious institutions or governments (Horváth, 2022). Each society can have its own moral system with a unique ethical identity (Bryson, 2018). Whether we view an action or decision as morally right or wrong depends on how we understand ethics. There are different views of ethics that are somehow related to each other within the broader field of moral philosophy. While this work focuses mainly on normative ethics, we also account for other views, such as meta-, descriptive, and applied ethics, as illustrated in Appendix I (Table 5).

Regarding *normative ethics*, we focus on the most common and well-known ethical theories when transferring the different views to the human-AI interaction context: deontology, consequentialism (utilitarianism being its best-known variant), and virtue ethics (Bankins & Formosa, 2023; Bilal et al., 2021). We added contractarianism and the theory of justice in Table 1 because some of the literature in our final sample refers to those ethical views.

Table 1 Ethical theories of normative ethics

Ethical Theories	Definition and Emphasis	Guiding Question	Foundational Philosophers
Deontology	Deontological ethics posits that actions are inherently right or wrong, irrespective of their consequences.	How should human-AI teams act to account for given norms, standards, and principles?	<i>Groundwork of The Metaphysic of Morals</i> by Immanuel Kant (1785), (Kant, 1991; Kant et al., 2012).
Consequentialism	Consequentialism suggests that the morality of an action should be judged by its consequences.	How should human-AI teams act to generate the greatest benefits for the greatest number of people?	<i>Utilitarianism</i> by John Stuart Mill (1861), (Mill et al., 1989); <i>An Introduction to the Principles of Morals and Legislation</i> by Jeremy Bentham (1988), (Bentham, 1988); (Höffe, 2013).
Virtue ethics	Virtue ethics emphasizes the role of individual character and inherent ethical motivation when weighing different maxims and deciding on moral behavior.	What kind of human-AI teams do we want to be/create?	<i>Nicomachean Ethics</i> (c. 340 BCE) by Aristotle (1962); (Simpson, 1992).
Contractarianism	Contractarianism argues that moral and political obligations arise from a social contract agreed upon by individuals for their mutual benefit.	How can human-AI teams operate based on agreements or contracts that ensure fair and mutually beneficial interactions?	<i>Leviathan</i> by Thomas Hobbes (1651), (Malcolm, 1996); <i>The Social Contract</i> by Jean-Jacques Rousseau (1762), (Rousseau, 1964).
Theory of Justice	The theory of justice focuses on the principles of fairness, particularly in the distribution of wealth, opportunities, and social benefits.	How can human-AI interactions be designed to ensure justice and fairness for all stakeholders involved?	<i>A Theory of Justice</i> by John Rawls (1971), (Rawls, 1971).

Deontological ethics is rooted in duty and obligation and posits that certain actions are inherently moral or immoral, regardless of their outcomes (Gaus, 2001; Unver, 2023). Immanuel Kant, a pivotal figure in deontological ethics, introduced the concept of the “categorical imperative,” which asserts that an action is morally right if it can be universalized without contradiction (Kant, 1991). With the evolution of AI’s abilities, there is a growing philosophical discussion regarding whether and how Kant’s categorical imperative can be applied to AI: some argue that AI should adhere to the law and follow previously defined ethical guidelines as objectives; others have a more subjective view of the imperative, arguing for using AI’s learning abilities for ethical reasoning by design (Dyrkolbotn et al., 2018). Regulations, such as the EU AI Act, can be seen as a form of deontological ethics (Koniakou, 2023). In work managed by algorithms, in particular, programmed carrot-and-stick mechanisms should guide ethical behavior (Millar, 2016; Wang et al., 2020).

Consequentialism determines the morality of an action based on its outcomes or consequences rather than, as in deontological ethics, some inherent rightness or wrongness (Yu et al., 2018). While utilitarianism is the best-known variant of consequentialism focusing on maximizing overall happiness or pleasure (Anderson & Anderson, 2007), other forms also exist, such as outcome-based morality, altruism, and egoism (e.g., Kago & Venkataraman, 2023; Wight, 2014), each prioritizing different types of outcomes. Utilitarianism posits that the morality of an action is determined by the extent to which it Jeremy Bentham and John Stuart Mill, prominent utilitarian philosophers, emphasized the “greatest happiness principle,” suggesting that actions should be evaluated based on their ability to generate the greatest good for the greatest number (Constantinescu et al., 2021, p. 806). Transferring this to human-AI interaction, a consequentialist view on morality means that humans interacting with AI, such as operating AI-based robots, may be judged, to some extent, regarding any negative outcomes because they have some level of responsibility to use AI for achieving a greater good (Simpson, 1992). Bryson (2018) compares this with the responsibility humans have for their pets or children and notes that the extent of sanctions for bad outcomes of human-AI interaction depends on how society measures the benefits of other options. Building “utilitarian” moral AI technologies would mean “reducing morality to a cost-benefit analysis” and requires thinking about whether being happy “can be unquestionably examined as a matter of goodness” (Serafimova, 2020, p. 6).

The concept of inner free will – i.e., humans can choose whether they behave ethically – is central to *virtue ethics* (Eitel-Porter, 2021; Flathmann et al., 2021). Hence, the moral intention behind an action or decision can guide moral actions and decisions, whereby AI technologies cannot have inherent virtue ethics (Heyder et al., 2023). Virtue ethics, rooted in the works of Aristotle, emphasizes the importance of cultivating virtuous character traits, such as courage and honesty, and posits that moral actions stem from having a virtuous character (Coeckelbergh, 2020; Hagendorff, 2020). These virtues can evolve over time into certain norms and guidelines, intertwining with deontological ethics, where they form duties (Eitel-Porter, 2021). The self-learning abilities of AI, based on human users’ data input and feedback, could be a way to integrate some form of virtue ethics into the machine in a collective form (Etzioni & Etzioni, 2017). Since AI lacks inherent virtues, this approach challenges us to consider how AI systems can be structured to promote ethical behavior in human users (e.g., Constantinescu et al., 2021; Neubert & Montañez, 2020).

Additionally, *contractarianism* focuses on ensuring fair and mutually beneficial interactions between human-AI teams through agreements or contracts, while the *theory of justice* guides the design of human-AI interactions to ensure fairness for all stakeholders (e.g., Coeckelbergh, 2009; Gabriel, 2022; Rafanelli, 2022).

The examples mentioned previously illustrate that each ethical theory offers different insights into principles and maxims that might be conceived and their application to specific situations. The selection and application of these theories by researchers can shape our understanding of morality within human-AI interaction significantly. The social contract that individuals and society agree upon (contractarianism), leading, in the best case, to increased fairness and wealth, following the theory of justice (Coeckelbergh, 2009; Erman & Furendal, 2022; Gabriel, 2022; Rafanelli, 2022).

To sum up, the divergence between different ethical theories plays a pivotal role in AI management, development, deployment, regulation, education, and usage (Brendel et al., 2021; Mirbabaie et al., 2022). Understanding these distinctions, as well as the interdependencies of different ethical theories, is essential for

navigating the ethical landscape of AI ethics. Without some combination of ethical views and knowledge of their limitations, there can be “bitter ideological battles, research agencies will become bogged down in disputes, and developers will split up their resources and devolve into a competitive mindset which reduces information sharing and slows research progress, thus making it more difficult to construct moral machines” (Bogosian, 2017, p. 606).

2.2 Transformative Nature of Human vs. AI’s Agency: Impact on Perceptions of Morality and Ethics

Beyond these ethical theories that govern our understanding of morality, it is essential to distinguish between the concepts of moral and ethical agency and agents. These terms have been used inconsistently in prior literature, leading to varied interpretations. This includes differing views on the traditional human moral agency (HMA) versus what is often called “artificial moral agency” (AMA) (Behdadi & Munthe, 2020, p. 195), “artificial moral agents (AMAs)” (Mabaso, 2021, p. 137), and the distinctions between “implicit” and “explicit ethical agents” (Moor, 2006, p. 19).

In general, the agency is “what’s going on inside” an agent (such as a human teammate or AI technology), i.e., an agent’s intentions, reasons, and maxims forming internal processes for making decisions and actions (Gogoshin, 2021, p. 2). Ontological frameworks, thereby, play a crucial role in defining humans and AI’s agency as different environments and artifacts create different realities, which impact our understanding of what ethical standards and moral values might be deemed appropriate (Cunneen et al., 2019; Orlikowski, 2010). Hence, ontological frameworks define agency and shape our perceptions of what is right and wrong, creating various “realities” regarding ethical AI design, deployment, and usage. Relational and substantial ontology are typical of the ontological frameworks (Cecez-Kecmanovic et al., 2014; Orlikowski, 2010). In the context of ethical human-AI interaction, relational ontology would suggest that the ethical behavior of both humans and AI systems is shaped through the interaction (Faulkner & Runde, 2010; Heyder et al., 2023; Mutch, 2013). By comparison, substantial ontology would treat AI technologies and humans as distinct entities with inherent virtuous characteristics and ethical attitudes that do not change based on their interactions (Faulkner & Runde, 2010; Heyder et al., 2023; Mutch, 2013). Thus, different ontological frameworks can lead to different perceptions regarding ethical obligations and moral maxims of humans and AI agents. Politicians, for instance, may emphasize regulatory frameworks based on deontological principles to prevent potential user misuse of AI, perceiving AI as a tool and its responsibility to be a human matter (relational ontology), whereas developers may prioritize technical safety to avoid privacy breaches as part of their virtue ethics maxims, perceiving AI as a (semi-)autonomous system (substantial ontology) (Cunneen et al., 2019; Khan et al., 2022; Kieslich et al., 2022). As AI technologies evolve and their capabilities advance, the ontological understanding of moral and ethical agency within human-AI interaction is crucial for ethical analysis (Cunneen et al., 2019).

Ethical agency involves reflecting upon what is “right” and “wrong” by weighing between different moral maxims against the often invisible background of one’s belief system and ethical assumptions, which can be interpreted through various ethical theories, as has been introduced in Section 2.1. Although the intention may be to identify the morally best option, it is crucial to recognize that not all decisions involve deliberate considerations of all options; many decisions are made habitually, potentially overlooking key factors. As a result, even well-intentioned decisions can lead to negative consequences. Well-intentioned AI, for instance, might ponder ethical decisions extensively, yet still produce poor outcomes if a critical factor is overlooked (Berber & Srećković, 2024; Champagne & Tonkens, 2023). Such poor consequences, however, can lead to corrections that yield significant

long-term benefits (Sætra & Danaher, 2023). By contrast, moral agency concerns perceptions of “good” and “bad” based on societal moral values and cultural norms, leading to recognition and accountability within a specific “moral community” (Champagne & Tonkens, 2023, p. 27:8; Mabaso, 2021; Søvik, 2022) and taking responsibility for one’s own actions and decisions (Behdadi & Munthe, 2020).

Agency is a prerequisite for acting responsibly, while responsibility encompasses the entire process from the conception, planning, and execution of an action to the evaluation of its outcomes. Responsibility is often understood as “the obligation to satisfactorily perform a task,” and accountability is “the liability to ensure that it is satisfactorily done” (McGrath & Whitty, 2018, p. 1). However, it should be noted that there is no common understanding of responsibility and accountability because these terms are often used as interchangeable synonyms (Lindkvist & Llewellyn, 2003). In essence, responsibility is about the “what,” referring to the duties and tasks assigned to individuals, teams, or organizations in managing, developing, and using AI technologies in a responsible manner. Hence, when talking about moral responsibility, it is about how people or organizations are capable and willing to take, act, and be responsible, referring to their agency and thus, to their inherent normative perceptions of what is right or wrong (Novelli et al., 2023). By comparison, accountability pertains to “who” is answerable for outcomes of human-AI interaction, which is based mainly on normative perceptions of society (Horneber & Laumer, 2023; Novelli et al., 2023).

Hence, “[w]hen humans are acting and making decisions, agency is normally connected with responsibility” (Coeckelbergh, 2020, p. 2052). Thus, the attribution of responsibility within human-AI teamwork is highly connected to morality (Passlack et al., 2023). Therefore, normative expectations, such as legal and social norms, can be explained through normative ethical theories that determine how we perceive responsibility and accountability. As Horneber and Laumer stated, even though organizations build AI technologies that give full control to users, they often remain responsible in terms of a “normative obligation to be accountable for these systems” (2023, p. 724) since their employees developed and operated the AI technology that demonstrates increasing levels of intelligence and autonomy. Following these arguments, designers developing and training AI and the organization providing and selling AI technologies are, to some extent, accountable for the ethical implications of AI utilization (Martin, 2019). To avoid responsibility gaps, we must be aware that with AI’s increased self-learning capabilities and human-like behavior (resulting in a novel form of agency), humans may perceive AI technologies “as moral agents that could act autonomously and be accountable for their behaviors” (X. Zhang et al., 2023, p. 1), which affects human-AI collaboration and could lead to uncertainties regarding responsibilities and accountabilities within the human-AI interaction (Figueras et al., 2022; Giermindl et al., 2022).

There are different scenarios in human-AI interactions to consider that shape our perceptions of humans and AI’s agency, such as humans directly interacting with AI systems, where the choice is limited to either using or not using the system (e.g., language translation tools), humans working in conjunction with AI agents’ outcomes, where humans have the option to follow or disregard the AI’s suggestions (e.g., decision-support systems), and AI actions without human involvement (e.g., autonomous driving). Following the ontological distinction of Barad’s (2003) agential realism, when humans and nonhumans (e.g., AI) are entangled as sociomaterial assemblages, a specific configuration of reality with particular moral values and norms determines the being and

knowing and, thus, how their actions are performed (referring to performativity as one of the tenets of sociomateriality; Orlikowski, 2007).

Different views on ethics can be used to illustrate how those actions are influenced. Humans decide within human-AI interaction how to respond to the technology based on their human agency (Cooren, 2004, p. 377; Leonardi, 2011), which is influenced by their ethical ideology and personal values, which can be linked to *virtue ethics*. By comparison, the AI's material agency can be described as "the way objects act when humans provoke it" (Leonardi, 2012, p. 37), which, for instance, can involve certain underlying norms that the AI follows by design that can be linked to *deontology*. Both human and AI agencies can also focus on beneficial outcomes – the AI via processing parameters and the human by choosing the maxima of achieving the greatest benefits, which can be linked to *consequentialism*. In cases of autonomously acting AI systems, we, thus, may perceive the AI's agency based mainly on the way it adheres to deontological principles, while in cases of supportive AI tools, we may account more for the user's agency (e.g., Niederman & Baker, 2023; Swanepoel, 2021), with its inherent moral norms and values (linked to virtue ethics). Additionally, in cases of training neural networks to identify cancer in X-rays, consequentialist maxims would be that the network arrives at decisions that can statistically outperform or, at least, complement human doctors, leading to the question of whether an AI ethical decision-maker needs to be perfect or simply better than humans? Hence, perceptions of human and material agency can be influenced through the type of entanglement between humans and AI that can shape their reality and ethical reflections along with societal values and moral maxims, which can be linked to different views on ethics.

Floridi and Sanders (2004) explore the ethical dimensions of artificial agents. They argue that artificial agents extend the traditional boundaries of moral involvement, as they can be considered both moral patients (entities that can be affected by actions) and moral agents (entities capable of performing actions that have moral implications). The authors suggest that morality can be understood as a "threshold" based on observable factors within a specific context or "level of abstraction" (Floridi & Sanders, 2004, p. 1). An agent is deemed morally good if all its actions adhere to this moral threshold and morally evil if any of its actions violate it. This idea challenges traditional notions of morality by including artificial entities as participants in moral situations, thereby expanding the scope of moral consideration to include nonhuman agents. Even though the emergent properties of the underlying models of some AI technologies can, to some extent, lead to the capability of ethical reasoning (Dyrkolbotn et al., 2018), the moral agency of humans is different from that of AI because the technology cannot act "freely" (as in 'having free will') or 'know' (as in 'being aware of') what it is doing" (Coeckelbergh, 2020, p. 2055).

This is why traditional HMA is different from AMA. The latter requires humans to exist (e.g., it arises through the moral values of developers setting certain AI parameters or using nondiscriminatory data to train AI models); it is also limited by HMA (Behdadi & Munthe, 2020). Moor used similar arguments as early as 2006, stating that AI is an "explicit ethical agent" that acts on behalf of ethical principles defined by humans and their HMA to make "judgments of duty" (2006, p. 20). In addition, AI might also be an "implicit ethical agent" that "act[s] ethically because its internal functions implicitly promote ethical behavior – or at least avoid unethical behavior [as] it has, to a limited extent, virtues" through learning from its human users' data input and feedback based on their underlying HMA (Moor, 2006, p. 19). Thus, when evaluating the agency of AI technologies, moral agency refers to the AI's capabilities to perform actions that align with the societal moral values and cultural norms of its

users and developers (i.e., ChatGPT learns to avoid generating responses deemed inappropriate in certain cultures based on user feedback), while the ethical agency of AI pertains to the system’s capacity to reflect on moral norms through its internal mechanisms, which are designed to promote ethical behavior and avoid unethical outcomes, such as ChatGPT’s refusal to generate content that aligns with extremist ideologies or other harmful requests (Krügel et al., 2023; Stahl & Eke, 2024). Thereby, in both scenarios, ChatGPT can also be seen as a moral patient because it is affected by the actions of developers and users. Vice versa, ChatGPT’s moral advice can influence users’ judgment, making users moral patients as well (Krügel et al., 2023; Stahl & Eke, 2024). Table 2 provides an overview of those somehow interrelated concepts.

Table 2 Definitions of ethics and morality in the context of humans and AI’s agency

Concept	Definition	Context: Human Beings	Context: AI Technologies	Intersections with Other Concepts
Moral Agency (Moral Agents)	Moral agency is the capacity of a moral agent to make decisions and act based on both societal moral values and cultural norms of a broader community, leading to perceptions of what is “good” and “bad.”	Human moral agency (HMA) comprises the autonomy of individuals as moral agents to either align with or diverge from societal moral values and cultural norms based on free will, subject to moral judgment for whether their actions are “good” or “bad” (Manna & Nath, 2021).	Artificial moral agency (AMA) is capable of executing tasks with moral implications more efficiently than humans in certain contexts (Behdadi & Munthe, 2020). However, they do not possess intent or free will. They operate under human-defined moral guidelines and parameters, reflecting the societal moral values and cultural norms of their human creators, making humans ultimately responsible for AI’s actions (Cervantes et al., 2020).	• As human values guide how they interact with AI technologies, their moral agency can be understood as a prerequisite to establishing ethical agency (one needs to have moral values to reflect on them) (Mabaso, 2021). • While AI technologies perform actions within the moral frameworks of its users or developers, AI is also a moral patient (Floridi & Sanders, 2004). • Unlike humans, AI lacks the capability to independently recognize or resolve ethical dilemmas (linked to ethical agency, which is the reflection of those dilemmas based on different ethical theories), making its form of moral agency fundamentally different (Behdadi & Munthe, 2020).
Ethical Agency (Ethical Agents)	Ethical agency involves the capacity of ethical agents to critically reflect on and evaluate what is “right” and “wrong” by recognizing and weighting between different moral maxims against the often invisible background of one’s belief systems that can be interpreted through the lens of various ethical theories.	The capacity for ethical agency in humans is a product of cognitive processes, emotional intelligence, and socialization, influenced by moral intuition and virtues and enabling humans to balance between moral maxims. Thus, not all decisions involve deliberate considerations of all possible moral maxims. Humans stand as full ethical agents, capable of understanding and adhering to ethical norms through virtue and reasoned judgment (Cheruvalath, 2023; Upton, 2009).	AI’s approach to ethical agency is driven by logical operations and programmed ethical principles, reflecting the intentions and objectives of its human developers or users rather than an inherent capacity for ethical reasoning. AI technologies can be categorized as implicit or explicit ethical agents. Implicit ethical agents exhibit a form of ethical reasoning derived from analyzing user data and feedback (Moor, 2006). Explicit ethical agents operate based on ethical guidelines explicitly programmed by humans without the capacity for independent ethical reasoning beyond those parameters (Moor, 2006).	• As a human’s ethical agency also involves knowledge of ethical norms and regulations to adhere to, it can be linked to deontology. Additionally, as the capacity to critically reflect on different moral maxims involves moral intuition and virtue, it is also related to virtue ethics (Bilal et al., 2021; Cheruvalath, 2023). • By comparison, AI’s ethical agency has, so far, been based on programmed rules and ethical principles. Thus, it can be linked mainly to deontology. As those rules are further given by their creators, AI technologies can also be seen as moral patients (Cheruvalath, 2023; Floridi & Sanders, 2004). • AI technology’s self-learning capabilities might enable them to analyze ethical reasons from the virtuous character of their human users based on human data input. Thus, AIs might be able to represent their user’s virtue ethics to some extent, thereby being also moral patients (Floridi & Sanders, 2004).

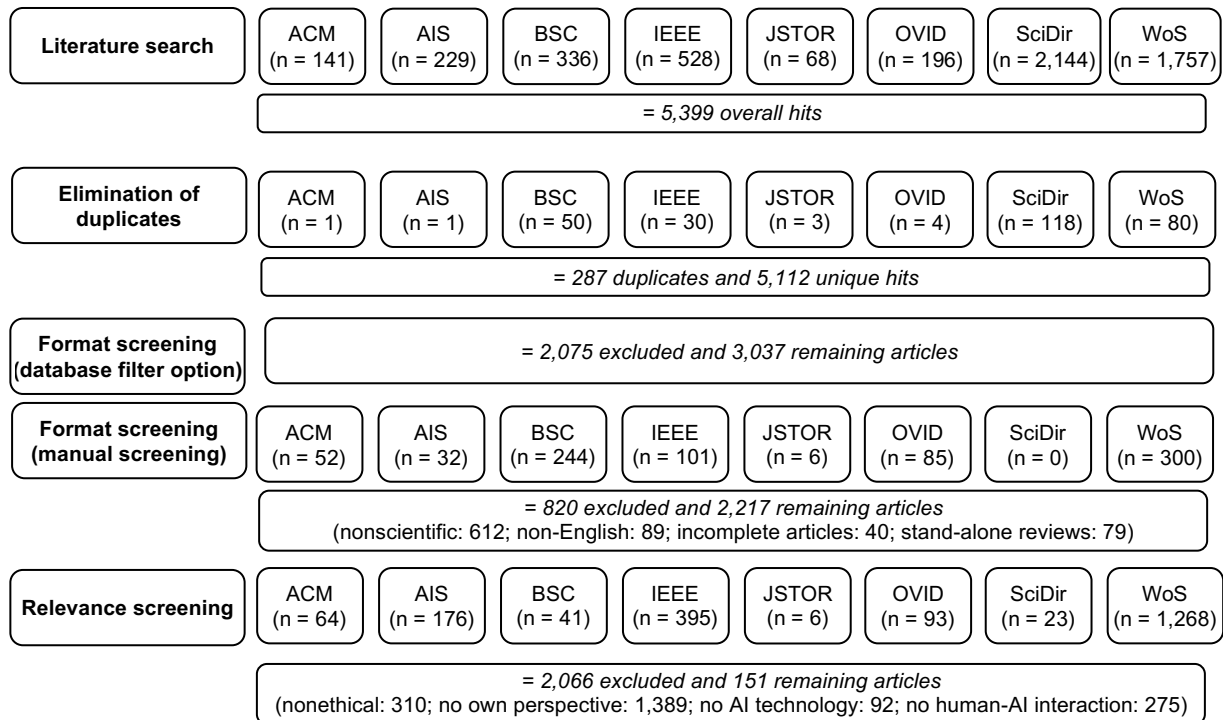
Moral Patency (Moral Patients)	Moral patency refers to the status of being an entity that is affected by actions to which moral considerations are owed, deserving ethical consideration and protection from harm.	Human beings are typically considered moral patients, as they are subject to moral consideration, deserving rights, protection, and ethical treatment regardless of their ability to make moral decisions (Floridi & Sanders, 2004; Shevlin, 2021).	AI technologies can be seen as moral patients in the sense that their treatment, programming, and usage reflect the moral values and ethical considerations of their creators. They are subject to ethical scrutiny based on the impacts of their actions on humans (Floridi & Sanders, 2004).	• Moral patency intersects with moral and ethical agency, as entities recognized as moral patients are granted moral consideration even if they lack agency (Shevlin, 2021).

We will use this background to guide the analysis of the literature review on how the moral and ethical agency of AI compared to humans is perceived, defined, and understood by the authors, as this highly impacts their conceptualizations on morality and responsibility in human-AI teamwork.

Method

Our method builds on a descriptive literature review that follows a systematic order. A descriptive review is appropriate to answer our research questions, as it helps us to identify underlying ethical theories of researchers and interpret certain trends regarding researchers' perceptions of moral agency (Paré et al., 2015). The systematic nature, following the steps of Templier and Paré (2018), further increases the transparency. Figure 1 shows the search process.

Fig. 1 Search process



Given that ethics-related research on human-AI interaction is scattered among different research disciplines, we searched the following eight databases to ensure broad coverage of different academic research areas: ACM Digital Library (ACM), AIS Electronic Library (AIS), Business Source Complete via EBSCOhost (BSC), also involving The Philosopher's Index for publications in the field of philosophy, IEEE Explore (IEEE), JSTOR, OVID, Science Direct (SciDir), and Web of Science (WoS).

We searched for relevant literature by querying the title, abstracts, and keywords – if available – for the keywords based on our theoretical background. This results in the following search string: (“ethic*/ethics” OR “moral*/morality” OR “ethical theory” OR “moral theory” OR “ethical philosophy” OR “moral philosophy” OR “consequentialism” OR “virtue ethics” OR “utilitarianism” OR “deontology”) AND (“Artificial Intelligence” OR “human-AI interaction”). The search, conducted from September to October 2023, yielded 5,399 hits overall. After eliminating any duplicates, we yielded 5,112 unique hits. Given the actuality of the literature identified and its broad coverage through the search in eight databases, we did not conduct backward and forward searches. The search string was carefully proven with a word cloud comparison of the main keywords arising in the citation network of n= 6,523 articles that are cited by our final sample, based on our bibliometric word analysis of the most often cited literature that yielded similar terms (see Appendix II, Figure 3-5).

Regarding *format screening*, we applied format criteria by eliminating nonscientific sources (e.g., gray literature), non-English publications, incomplete articles (e.g., extended abstracts, research-in-progress articles), and standalone literature reviews that do not contribute further to theory (e.g., reviews that merely aggregate previous findings without offering new insights, extensions, or developments to existing conceptualizations, frameworks, or theories). We did not restrict our search to a specific time range or certain quality criteria (e.g., rankings, peer-reviewed literature). If available, we used database filter options for format screening before screening manually (see Appendix III, Table 6). This results in the elimination of 2,075 articles through database filter options, 820 articles through manual screening (total: 2,895 articles), and 2,217 articles remained for relevance screening.

Regarding *relevance screening*, we applied additional criteria to ensure the alignment of articles with our research focus (see Appendix IV, Table 7, for details). We excluded articles that lack a connection to different views on ethics, particularly those that do not offer any ethical theory or philosophical perspective. Furthermore, we removed articles that fail to present an original viewpoint on ethics and morality in the context of human-AI interaction from the theorist’s or researcher’s standpoint, as these articles do not contribute directly to our research topic. We also excluded studies that do not address technologies aligning with AI’s evolving agency, specifically those demonstrating unique human-like, self-learning, and (semi-)autonomous capabilities. Lastly, articles that do not explicitly focus on human-AI interaction, such as those that only examine the functionalities of autonomous AI without considering the human element, were excluded. A total of 2,066 articles were excluded, and 151 articles remained as our final sample. Table 8 in Appendix V provides an overview of the final sample. It summarizes various perspectives on ethical human-AI interaction, morality, and agency, as well as the type and applications of ethical theories and the contextual and methodological approaches of each article. Following MacInnis’ (2011) idea of explicating types of conceptual contribution, Table 8 (Appendix V) aims to distill relevant contributions from each article that are most relevant to our research project through reduction and focus. This synthesis is based on our qualitative content analysis which is detailed below.

Regarding *qualitative analysis*, the articles of our final sample were downloaded as PDFs and analyzed through manual coding using Microsoft Excel. Firstly, we structured the samples’ meta-information (e.g., regarding the research domain, method, and ranking of the publication outlet) to provide a descriptive overview of the different article types within the final sample (see Appendix VI, Table 9). We then opted for qualitative analysis to organize the content of relevant articles systematically. We initially employed a deductive approach to structure the

findings, using the theoretical background as a guide (Mayring, 2014). In addition, we incorporated inductive codes, allowing us to include new observations that emerged during the analysis. This approach involved applying pre-established codes based on our theoretical framework while also adding inductive codes for insights not covered by the initial schema. The coding schema is detailed in Appendix VII (Table 10). Each coding category includes a name, an explanation, its underlying codes, the coding rules, and an anchor example from the final sample, as shown in Table 10. The coding process follows Mayring’s (2014) steps: *firstly*, define the research question and background; *secondly*, establish the category system (deductive); *thirdly*, define coding guidelines; *fourthly*, conduct a preliminary material run-through with coding and rules (adding initial inductive codes); *fifthly*, revise categories, codes, and rules; *sixthly*, working through the material with several rounds and discussion among coders; and, *finally*, analyze and interpret the coding. In order to enhance the quality of our coding process and redefine the coding schema, we used 62 articles for parallel training purposes of the coders based on the coding schema. We measured interrater reliability using Cohen’s Kappa, following Cohen (1960), to assess the agreement on all deductive codes, achieving a score of $\alpha=0.72$ in the first round and $\alpha = 0.79$ in the second round, which indicates a substantial agreement between coders (Burema et al., 2023). Concerning the inductive codes, which introduced novel categories, we engaged in extensive discussions among coders and iteratively redefined each code over rounds of coding.

We additionally applied a *citation network analysis* to map the intellectual landscape and reveal how different ethical theories are interconnected across various disciplines. This allows us to identify the most influential works, trace the evolution of key ideas and emerging ethical frameworks (contributing to RQ1), and detect ethical blind spots in ethical AI research as underrepresented or isolated clusters in the network which may indicate areas where certain ethical perspectives are overlooked (contributing to RQ2) and helps us to understand how AI transformation challenges traditional notions of moral agency and responsibility by highlighting the integration (or lack thereof) of new ethical considerations into established discourses (contributing to RQ3). We employed bibliometric packages in R for this analysis to examine publication dynamics based on different bibliometric analysis techniques following Zupic and Čater (2015); see Table 3. We refined our dataset by extracting relevant metadata for the articles from the Dimensions.ai database, which is well-suited for analysis with the bibliometric R package. Metadata for 137 articles was successfully gathered from a total of 151 articles considered, resulting in a total number of 6,523 analyzed articles (“global citations”) involving the references of those 137 articles (see Table 11 in Appendix VIII for details).

Table 3 Overview of bibliometric analysis techniques

Method	Motivation	Description	Unit of Analysis
Citation analysis	Estimates the influence of documents, authors, or journals through citation rates.	Citations are used as a measure of the influence of the top-N list of the studies, authors, or journals most cited. Thus, if an article is heavily cited, it is considered important (Zupic & Čater, 2015). While citation analysis offers insights into the publication’s relative impact, it does not capture the interconnected networks among scholars (Üsdiken & Pasadeos, 1995).	• Document • Author • Journal
Co-citation analysis	Connects documents, authors, or journals on the basis of joint appearances in reference lists.	The co-citation analysis was used to measure the similarity between documents, authors, and journals based on how frequently they are cited together (McCain, 1990; Small, 1973). This method enables the identification of networks and relationships within the academic landscape, reflecting the scholarly community’s valuation and interaction with research over time.	• Document • Author • Journal
Bibliographic coupling	Connects documents, authors, or journals on the basis of the number of shared references.	Bibliographic coupling was used to quantify the similarity between two documents based on the number of shared references in their bibliographies, with a greater overlap indicating a stronger connection (Kessler, 1963). This ensured that bibliographic coupling was constrained within a given timeframe to ensure coherence (Glänzel & Thijs, 2004).	• Document • Author • Journal

Co-author analysis	Connects authors when they co-author the paper.	Co-author analysis was used to demonstrate relationships between co-authors (Acedo et al., 2006; Lu & Wolfram, 2012). When two documents frequently appear together in citations, it implies that each document also garners a high number of citations independently (Jarneving, 2005). We applied the Fruchterman-Reingold visualization technique and the Edge Betweenness clustering algorithm (repulsion force 0.1; minimum number of edges = 2).	• Author
Co-word analysis	Connects keywords when they appear in the same title, abstract, or keyword list.	Co-word analysis was used to map conceptual frameworks within a field based on the title, keywords, abstract, or full text (Callon et al., 1991).	• Word

Results

In this section, we answer our three research questions. Section 4.1 provides insights into the fragmented ethical streams that exist in current AI ethics research (corresponding to RQ1); Section 4.2 elaborates on challenges that emerge when establishing ethical frameworks (corresponding to RQ2); and Section 4.3 discusses how AI advancements challenge traditional notions of moral agency and responsibility (corresponding to RQ3).

4.1 The Role of Interdisciplinary and Global AI Ethics Perspectives

This section addresses our first RQ: “*What ethical perspectives are addressed in AI ethics research?*” Below, we illustrate how an interdisciplinary and international academic discourse would contribute to integrating diverse ethical perspectives within AI ethics research. Table 4 provides an overview of the distribution of normative theories within AI ethics research, outlining that 49 articles focus on specific ethical views while others combine them:

Table 4 Distribution of normative theories (n = 93)

	n (= 93)	%	Example from the final sample
Deontology	21	22.58 %	Agbese et al. (2023) or Behdadi & Munthe (2020)
Virtue Ethics	11	11.83 %	Bilal et al. (2021) or Boyles (2018)
Consequentialism (Utilitarianism)	7	7.53 %	Emdad et al. (2023) or Gerdes (2022)
Contractarianism	10	10.75 %	Buruk et al. (2020) or Carneiro & Veloso (2022)
All three most common theories (deontology, virtue ethics, and consequentialism/utilitarianism)	16	17.20 %	Génova et al. (2023) or Telkamp & Anderson (2022)
Deontology and virtue ethics	19	20.43 %	Bankins & Formosa (2023) or Bogosian (2017)
Deontology and consequentialism	5	5.38 %	Colmenarejo et al. (2022) or Corvite et al. (2023)
Deontology and contractarianism	4	4.30 %	Ashrafian (2015a) or Dodig Crnković & Cürüklü (2012)

How the AI ethics discourse can be fragmented into two main streams

The citation network analysis (see Appendix IX, Figure 6) reveals two distinct clusters of academic discourse centered around ethical principles in AI. The first cluster builds upon the work of Jobin et al., who outline a “global landscape of AI ethics guidelines” (2019) and the five critical “AI4People” principles of Floridi (2018, p. 689): beneficence, non-maleficence, autonomy, justice/fairness, and accountability. This cluster, which includes work by Hagendorff (2020), Hawkins and Mittelstadt (2023), Martin (2019), and others, emphasize the design and regulation of AI technologies closely linked to deontology. Specifically, 69.89 % of the articles in this cluster address AI ethics by referencing these deontological principles, often prioritizing the regulation of AI systems based on established rules and guidelines to “morally regulate the use of AI” (Pansoni et al., 2023, p. 1152). This reflects a top-down approach, where ethical behavior is enforced through rules and regulations imposed by higher authorities, such as governments or regulatory bodies. However, the legitimacy of these authorities being responsible for establishing and enforcing ethical rules can vary widely between nations and might not always be based on democratic processes (H. Roberts et al., 2023; Tasioulas, 2018). For instance, there can be nations where

power is concentrated in the hands of a few, wherein the ethical frameworks imposed may reflect the interests of the ruling elite rather than a broader societal consensus. This can lead to regulations that are rigid, narrow, and potentially oppressive, failing to account for diverse perspectives and values (H. Roberts et al., 2023). By predominantly framing AI ethics through the lens of deontological ethics, other ethical views – such as virtue ethics and consequentialism – may be underexplored.

The second cluster in our citation network revolves around the works of Allen et al. (2000) and Coeckelbergh (2020), following the idea of human-centered AI (HCAI). This group of articles engages with AI ethics at a more conceptual level, focusing on virtue ethics and consequentialism to promote social justice through responsible AI design, implementation, usage, and monitoring. This cluster, which traces its discourse back to 1985 with Lloyd (Lloyd, 1985), gained significant momentum in 2019 (see the peak in Appendix X, Figures 7-8), highlighting the broader implications of AI on ethics. The notion that AI systems, as active participants in human environments, should adhere to human values and principles is central to HCAI (Ghatar et al., 2023; Lloyd, 1985; Millar, 2016; Nevanperä et al., 2021). This includes “ensuring transparency, fairness, equality, explainability and avoiding bias or discrimination” in a human-centered approach (Carneiro & Veloso, 2022, p. 1). Thereby, HCAI-related research in our reviewed articles often aligns with virtue ethics and consequentialism. This perspective emphasizes reducing the downsides of widespread AI use and maximizing its benefits for humans (Allen et al., 2000). However, only a small fraction of the articles in our sample (7.95 %) address how to guide humans during the implementation of AI solutions from a human resource management (HRM) viewpoint (e.g., Martin, 2019; Rodgers et al., 2023). Thus, these works risk underestimating the real-world dynamics that influence AI ethics by not grounding their discussions in empirical evidence, which could lead to an incomplete understanding of how ethical principles are applied in practice (e.g., Bleher & Braun, 2023; Townsend et al., 2022). This raises the question of whether the discourse within this cluster adequately captures the complexities of implementing ethical AI in real-world contexts, where institutions and interest groups play a significant role. Bridging the gap between conceptual discussions and empirical research could strengthen this cluster’s contribution to the broader AI ethics discourse (e.g., Bleher & Braun, 2023).

To summarize, these varying emphases within the two clusters outline the relevance of interdisciplinary collaboration between different research streams. Therefore, Unver (2023) recommends collaboration between industries, universities, and tech partners to revise lawyers’ rules and industrial codes of conduct. Combining the strengths of the different clusters would ensure a comprehensive approach to AI ethics.

How the AI ethics discourse outlines geographical contours

Our bibliometric analysis of the citation network highlights the evolving ethical trajectory in AI research and its geographical contours. There has been a significant growth in the infusion of ethical considerations into AI scholarship since 2019, particularly within the IS field. However, this growth is not uniform across the globe. Research is disproportionately clustered in regions such as North America, Europe, and Asia-Pacific, revealing strong geographic disparities (J. S. Roberts & Montoya, 2022). This spread of ethical discourse suggests that the ethical questions being asked – and the answers being proposed – are shaped by the specific contexts and priorities of different regions (H. Roberts et al., 2023; Stahl et al., 2023). Exemplarily, in some countries, AI not yet have penetrated markets deeply and may be seen primarily as a black box component in systems developed elsewhere

(Polyviou & Zamani, 2022; H. Roberts et al., 2023). This limits the scope of ethical discourse and highlights the importance of fostering greater global participation.

When discussing international collaboration, it is essential to recognize that AI's presence, applications, and societal impact can vary significantly across different regions. The extent to which AI technologies are integrated into daily life, the sectors in which AI is predominantly used, and the cultural attitudes toward AI differ from one country to another (H. Roberts et al., 2023; Wu et al., 2020). These variations mean that ethical concerns and priorities in AI may also differ and be shaped by local contexts and cultural perspectives. Our definition of international collaboration considers partnerships between researchers from different countries, regardless of their personal or ethnic backgrounds. This underscores that geographic and institutional diversity is essential for broadening AI ethics research (Stahl et al., 2023). Incorporating diverse perspectives enables researchers to better address the ethical challenges that vary across regions. The need for such collaboration is evident in the distinct author networks that tend to cluster around specific regions – such as the United States and Canada, Europe, and Asia-Pacific – rather than crossing international boundaries (see Appendix XI). This segregation suggests that the dialogue on AI ethics is heavily influenced by regional regulatory frameworks of dominant nations, such as the U.S. Bill of Rights, the EU AI Act (e.g., Colmenarejo et al., 2022; Stahl et al., 2022), and the Beijing AI principles for the Beijing Academy of Artificial Intelligence under China's Ministry of Science and Technology (e.g., Wu et al., 2020). Thus, dominant nations, such as the USA or China, often set a certain focus on regulation, which is then reflected in academic discourse because the United States and China lead in AI research volume (e.g., Telkamp & Anderson, 2022). While in Europe, for example, ethical principles such as transparency and privacy are strongly promoted by the EU AI Act and the General Data Protection Regulation (GDPR), China tends to pursue state security with cybersecurity law (e.g., Stahl et al., 2022; Wu et al., 2020). In doing so, the EU strives to protect individual users following deontological maxims, while China tries to protect its institutions and nation following a utilitarian maxim. These frameworks driven by dominant nations are important in shaping local AI ethics. However, they may not fully account for the diverse ways in which AI is experienced and understood globally. This finding underscores the critical role that diverse scholarly perspectives play in steering responsible AI policies. International collaboration and interdisciplinary dialogue are essential to address these blind spots and foster a more comprehensive ethical discourse.

In sum, while international collaboration enriches AI ethics research, it is the accumulation of diverse regional and disciplinary perspectives that truly addresses AI's complex ethical challenges. Thus, fostering international co-authorships can be an important strategy for expanding the reach and impact of AI ethics research since researchers play a crucial role in shaping responsible AI policies (Rismani & Moon, 2023). However, we still acknowledge that the necessity of international collaboration varies depending on the research focus (see Appendix XII). Some technical guidelines may require less global input, while normative and culturally specific issues benefit greatly from it. Thus, international partnerships involving diverse perspectives are crucial to account for the complex nature of AI ethics research.

4.2 Challenges in establishing ethical frameworks

This section deals with several challenges in establishing ethical frameworks for human-AI interaction, particularly corresponding to the second RQ: *“What challenges for establishing ethical frameworks are emerging within AI ethics research?”* While we highlight the challenges in establishing ethical AI frameworks, our findings

suggest that no single, universally applicable ethical AI framework exists. Instead, the suitability of any framework depends on the specific context and use case, as ethical considerations are shaped by varying societal values, cultural norms, and application domains.

Should frameworks be tailored to the specific AI functions, or is it feasible to regulate AI as a unified entity?

Our review highlights that a lot of the current research on AI regulation, mainly linked to deontology, tends to address AI only on the meta-level without adequately accounting for the heterogeneous nature of AI technologies, their underlying models, processing mechanisms, and application domains. Wright (2020) argues that many ethical guidelines are on a higher level; thus, specific guidelines are required in particular domains to identify and assess ethical risks. The prevalent assumption that AI can be regulated as a singular entity oversimplifies the complexities inherent in various types of AI applications (Wright, 2020). Applying the “do no harm” principle to facial recognition technology, for instance, involves considerations of privacy and potential biases that could lead to wrongful identification or discrimination (e.g., Tasioulas, 2018; Zebrowski, 2021). By contrast, the same principle for translation tools might focus on the accuracy of language processing and the ethical implications of mistranslations, which could result in miscommunication or cultural misunderstandings (e.g., Greene et al., 2019). Similarly, autonomous weapons systems raise entirely different regulations for the “do no harm” principle, particularly concerning potential unintended harm to civilians (e.g., Gómez De Ágreda, 2020). Thus, different types of AI have conflicting moral foundations (e.g., Telkamp & Anderson, 2022).

The unique functions and impacts of each technology make it challenging to establish a universal regulatory framework based on a single set of principles, norms, or guidelines, further enforced by carrot-and-stick mechanisms regarding compliance or violation (Swanepoel, 2021; Wang et al., 2020). Fabiano (2019) highlights that ethics is highly based on the implicit knowledge and activities of the individual operators. Thus, the environment, programming, and using AI technologies mainly determine the agency of AI (Gahnberg, 2021). Therefore, frameworks of AI governance should be aware of this environment, forming expectations of AI agents, and calling for specific regulations (Gahnberg, 2021). Additionally, Mäntymäki et al. (2023) criticized that often frameworks of AI governance are based on inconclusive findings of the research, focusing on specific situations with specific AI-related risks. The failure, for instance, to adhere to transparency principles in conversational agents, such as ChatGPT, where the primary concern is ensuring users are aware they are interacting with AI, may have vastly different implications than noncompliance with transparency norms in AI-driven X-ray examinations (e.g., Burema et al., 2023; Lloyd, 1985). In the latter case, clarity of diagnostic outputs and the explainability of medical conclusions are crucial for both medical professionals and patients. This variation in the application and impact of transparency principles underscores the difficulty of creating a one-size-fits-all regulatory framework or the establishment of a single standard for AI (Jantunen et al., 2021).

These examples underline the challenges inherent in formulating a comprehensive regulatory framework for AI that follows deontological ethics. While broad principles such as “do no harm” provide a starting point, they may offer limited guidance if the specific harms associated with various AI technologies are not thoroughly understood or addressed (Wright, 2020). The diverse range of AI applications necessitates a more nuanced approach to regulation, one that recognizes the unique risks and ethical considerations of each technology (Jantunen et al., 2021). This complexity is further amplified by the various interpretations of AI’s agency, which

can impact allocations of responsibility (Pansoni et al., 2023; J. Zhang & Zhang, 2023). If AI, for example, is viewed as subordinated to its designer's moral proxies programming how the system should operate or the moral conduct of a human user operating assistive tools, it would be considered as a moral patient, dependent on the designer's or user's decisions as a moral agent (Sun & Ye, 2023). Conversely, if AI is perceived as being capable of moral action through its self-learning mechanisms and autonomous behavior, where neither the developer nor the user can see exactly how the AI is processing to come to a decision, AI might be held accountable as a moral agent (Sullivan & Fosso Wamba, 2022). In this context, understanding an AI's moral intention, or the ethical considerations behind its decisions, can be addressed by embedding moral principles directly into its programming (Badea & Artus, 2022). This approach allows developers to guide the AI's decision-making process toward ethical outcomes, thereby shaping the AI's actions according to predefined moral standards. Thus, the specific characteristics of AI technology lead to varying perceptions of moral agency, which, in turn, affect how responsibilities are allocated.

The need for clarity is particularly pronounced when considering the prosecution of non-compliant behaviors, such as criminal decisions, arising from human-AI interaction (Tajalli, 2021). It is essential to delineate the extent of human involvement in these situations (Tajalli, 2021). However, as noted by a criminal law attorney in a study by Walke et al., "there are hardly any concrete AI use cases in the area of law enforcement [...]" (2023, p. 3). The challenge is compounded by the multitude of stakeholders who could be held accountable, given that AI decisions and actions are often influenced, directly or indirectly, by the morality of the human creators, maintainers, or users of these technologies (Shevlin, 2021; Swanepoel, 2021). Furthermore, the vast array of AI technologies – from autonomous systems to augmentative tools – complicates the issue. Some authors argue, for instance, that AI technologies are merely explicit ethical agents that respond only to the principles for which they were programmed, thereby maintaining that AI creators must always retain responsibility (e.g., J. Zhang & Zhang, 2023). Others contend that responsibility can also be attributed to employees as users making decisions and actions in conjunction with, or independent of, augmentative an AI system's suggestions, such as algorithmic assistants that provide advice (e.g., Coeckelbergh, 2020; Cooper et al., 2022).

Consequently, defining appropriate governance principles to guide AI design, usage, and management (specifically regarding legal responsibilities) is a complex task. Establishing an ethical baseline that accommodates the diverse interests of human stakeholders across the board is already challenging enough (Khan et al., 2022; Volkman & Gabriels, 2023). This difficulty is compounded when considering that these standards may need to be interpreted differently for various technologies. As Roberts et al. (2023, p. 91) state the "EU's experience from the High-Level Ethics Group could also act as a model for introducing baseline cross-sectoral guidance"; however, such attempts thus far "are extremely high-level and vague." The authors advocate for greater integration of public users from various domains and AI application areas in the formulation of laws and policies since lobbying by huge AI providers "lead[s] to a problematic watering down" of current regulations, such as the EU Act (H. Roberts et al., 2023, p. 92). The authors advocate for greater integration of public users from various domains and AI application areas in the formulation of laws and policies. This is crucial, as lobbying by large AI providers, particularly those with widespread business-to-customer applications, such as social media or generative AI, often leads to the dilution of regulations, such as the EU Act (H. Roberts et al., 2023). Without such involvement, there is a risk that dominant AI providers might shape regulations and create bottleneck situations, which can be challenging for machine learning projects in more niche domains, making AI regulation

unsuitable for more specialized business-to-business (B2B) AI applications from niche domains, such as breast cancer detection systems (e.g., Gerdes, 2022).

In order to address these challenges, it is essential to involve the public and users from diverse AI application domains in the regulatory process. This broader participation could lead to the development of ethical frameworks with principles that are tailored to the specific characteristics and risks of different AI technologies (Wright, 2020). By making collective moral deliberation an integral part of the “process of law- and regulation-making” (Etzioni & Etzioni, 2017, p. 405), AI governance can become both comprehensive and adaptable, and better equipped to meet the specific challenges posed by the wide array of AI technologies. This would require aiming to achieve consensus among AI researchers, executives, and policymakers on appropriate frameworks of global AI governance as a difficult international effort and an ongoing process rather than a final outcome (B. Zhang et al., 2021). However, this might also make it hard to find the minimum threshold for acceptable legal standards, potentially lagging behind the higher moral values of individuals. As Tóth et al. (2022, p. 8) noted that “Immoral is any action that only reaches the legal threshold’s bare minimum.” Thus, establishing frameworks might require hybrid approaches that consider tailored AI regulation, which not only addresses the unique characteristics of AI technologies and their impact but also aims to achieve a unified level of principles so that tailored frameworks do not potentially lag behind the moral perceptions of individuals.

Should frameworks prioritize the moral values of several communities, or is it feasible to prefer one moral value system?

Our review outlines that the rapid development and integration of AI into various aspects of human life have sparked an extensive debate on whether morality can ensure ethical human-AI interactions (e.g., Khan et al., 2022; Sullivan & Fosso Wamba, 2022). This debate is further complicated by the issue of whose morality is being referenced, as morality is not monolithic but is instead understood within diverse cultural, social, and individual contexts closely linked to virtue ethics (Champagne & Tonkens, 2023). Thus, the question arises: Whose moral framework(s) should guide ethical conduct in human-AI interaction?

The AI systems incorporate morality in the form of guidelines that enable them to operate ethically based on the moral perspectives of their creators, sticking to prevailing norms (e.g., Behdadi & Munthe, 2020; Génova et al., 2023). This approach is often aligned with deontological ethics. By contrast, human morality reflects a belief system regarding what constitutes ethical behavior, shaped by the cultural and social context in which individuals are embedded (Horváth, 2022). This can be linked to virtue ethics, which emphasizes character and moral development. Consequently, both human and AI moralities are tied to specific moral systems (Champagne & Tonkens, 2023). Our bibliometric analysis highlighted that the understanding of AI principles in Western cultures might differ significantly from that in Eastern cultures (Telkamp & Anderson, 2022). While AI technologies can be programmed to act in morally right ways, ethical conduct involves more than adhering to a specific code; it also requires sensitivity to the nuances of cultural and contextual variations (Génova et al., 2023). The clash of these moral perspectives could lead to ethical conflicts in AI decision-making processes, where the system’s actions may align with one moral framework but violate another. Thus, the question of whether morality can ensure ethical human-AI interaction is deeply intertwined with the issue of whose morality is being referenced.

Given the diversity of moral systems across and within countries, cultures, and contexts, relying on a single framework to guide AI is problematic. Instead, a more nuanced approach that considers multiple moral

perspectives that shape moral evaluation besides technically implemented principles in the AI design and seeks to find common ground may be necessary (Wehrens et al., 2023). This would require AI systems to be capable of understanding and integrating diverse moral principles, potentially through advanced machine-learning techniques that allow for ethical reasoning based on context (Floridi et al., 2018). Etzioni and Etzioni (2017), for instance, argue that AI's self-learning mechanisms could potentially develop collective ethical conduct based on diverse human users' data input and feedback having several perspectives of morality, which might be more objective than integrating a single set of principles from the developers into the AI's design. This would mean that the morality of different users guides ethical conduct besides that of those creating the AI. However, while this approach acknowledges the complexity of morality, it also raises concerns about the potential for moral relativism, where any action could be justified depending on the chosen moral framework (e.g., Conitzer et al., 2017; Serafimova, 2020).

The introduction of multiple perspectives on morality into human-AI interactions also increases the degree of uncertainty and potential biases, which may lead to moral relativism or even unethical outcomes (e.g., Serafimova, 2020). Conitzer et al. (2017, p. 4832) note that "When we code moral values into AI, we have the option of either using the moral values of a specific individual or group – a type of moral relativism – or giving the AI some type of social-choice-theoretic aggregate of the moral values that we have inferred (for example, by letting our models of multiple people's moral values vote over the relevant alternatives, or using only the moral values that are common to all of them)." Thus, implementing such a pluralistic approach presents significant challenges, both in terms of practical application and the risk of moral relativism. Volkman and Gabriel (2023, p. 8) state that "Pluralism itself is controversial, especially insofar as monist critiques associate it with some kind of moral relativism, but this only reinforces how great a challenge the diversity of moral opinion is for implementing moral knowledge in an AI system."

In conclusion, the question of whose morality should ensure ethical conduct in human-AI interaction cannot be answered by relying on a single framework. Given the diversity of moral systems across cultures and individuals with their underlying virtuous characters, it is clear that no one set of moral principles can universally govern ethical human-AI interaction (Champagne & Tonkens, 2023; Telkamp & Anderson, 2022). Instead, a pluralistic approach is necessary, one that considers and balances multiple moral perspectives in AI decision-making. This approach requires AI systems to be capable of integrating diverse moral perspectives, allowing for context-sensitive ethical reasoning, such as considering the specific ethical issues arising with the use of AI systems in healthcare (Buruk et al., 2020). However, this pluralistic approach also brings challenges, particularly the risk of moral relativism, where actions might be justified based on various ethical frameworks following different perceptions of virtue ethics (e.g., Conitzer et al., 2017). To explain this issue from a Habermasian perspective, the idea of translating morality to particular contexts might lead to the application of maxims in multiple ways; thus, prioritizing or choosing among them becomes imperative (Buhmann & Fieseler, 2023). Hence, establishing ethical frameworks might require context-sensitive and pluralistic approaches that also address the risks of moral relativism.

Should current AI capabilities or public perceptions of AI drive ethical frameworks?

Our review underscores the diverse perceptions and understandings people hold regarding the ethical or unethical behaviors of AI technologies compared to humans (e.g., Constantinescu et al., 2021; Hooker & Kim,

2018; Telkamp & Anderson, 2022). The discourse often centers on AI's potential evolution and the preemptive strategies needed to mitigate future ethical challenges rather than its current capabilities. At present, most of the literature reviewed considers AI technologies to be explicit ethical agents with limited moral agency, constrained by their programmed principles (e.g., Dyrkolbotn et al., 2018; Farisco et al., 2020; Tajalli, 2021). However, these studies also point to a “theory-practice gap” between the actual capabilities of AI technologies and the expectations held by various stakeholders (Bleher & Braun, 2023, p. 21:1). As Wong (2021) outlined, regulatory systems do not cover AI transformation sufficiently because they are created before the advances of AI technologies arise, which requires adaption. Also, Vakkuri et al. (2019, 2021) highlight the challenge of translating principles of AI ethics into real-world practices, and Hauer (2022, p. 1) criticizes that the concepts of ethics and ethical decision-making of AI technologies are often discussed too abstractly, not considering “particular physical organism in which the algorithm takes place, nor on what it is made of.” Thus, the question arises whether current AI capabilities or perceptions of future ones should drive the establishment of ethical frameworks.

One possible explanation for this gap lies in the ambivalent attitudes people have toward AI's potential for beneficence, with opinions often polarized toward enthusiasm and concern (e.g., Fukuda-Parr & Gibbons, 2021; Kučera, 2019). Interacting, for instance, with generative AI technologies or assistive tools that exhibit emotionality, human-likeness, or a semblance of ethical reasoning can lead to misconceptions about AI's abilities to replicate human behavior and morality (Bilal et al., 2021; Corvite et al., 2023; Zebrowski, 2021). Despite the fact that current conceptualizations and research on AI ethics view AI technologies mainly as explicit ethical agents (e.g., Dyrkolbotn et al., 2018; Farisco et al., 2020; Tajalli, 2021), participants in these studies often do not share this view. Moral agency for humans is shaped by the societal moral compass, legal frameworks, and educational systems that enhance their ability to critically reflect on ethical issues, which, in turn, influences their perceptions of AI and its morality (Constantinescu et al., 2021; Córdova & Vicari, 2022; Martinho et al., 2021). Key factors, such as the perceived degree of AI's autonomy, transparency, and explainability of AI, are crucial for humans to establish realistic perceptions of AMAs (Buhmann & Fieseler, 2023; Vanhée & Borit, 2022). Nevertheless, the relationship between AI's actual capabilities, its moral agency, and public perceptions is probably complex and interdependent. The risks, for instance, associated with AI, such as potential discrimination in job evaluation tools, can be “overshadowed” by perceived benefits, like increased efficiency in recruitment processes (Corvite et al., 2023, p. 1). Similarly, while the long-term benefits of AI might outweigh immediate ethical concerns for some, others may react strongly to even minor transgressions due to their fear or distrust of the technology (e.g., Chaput et al., 2021). This can lead to utopian or dystopian ethical narratives that limit the possibilities of enabling appropriate ethics for robots and AI (Tasioulas, 2018). Giroux et al. (2022, p. 1035) observed that “Indeed, when dealing with a human or a technology perceived as more human, people expect more guilt when thinking about a potential transgression of their moral standards.” Therefore, the gap between AI's current capabilities and moral agency and public perceptions can lead to varying requirements for AI ethics, resulting in different ethical frameworks.

Another factor contributing to this gap might be differences in conceptual and practical ethical issues. Conceptual issues involve the abstract and ontological nature of AI, as well as the aggregation of various interpretations of ethical theories. Terzis (2020, p. 244), for instance, stated that “[a]greeing on ethics for all is an illusion.” Champagne and Tonkens (2023) discuss metaethical fictionalism or instrumentalism in the context of autonomous robots, while Siapka (2022) offers a feminist metaethics perspective aimed at addressing gender-

related moral ignorance. By contrast, practical issues surrounding, among others, the human fear of being replaced by AI technologies in the labor market lead to either avoidance of AI or an overreliance on it due to the hype surrounding AI's abilities (Farisco et al., 2020). Thus, while conceptual issues concern the lack of consensus on AI ethics (Siau & Wang, 2020), practical concerns often stem from unrealistic perceptions of AI's capabilities, which are shaped by utopian or dystopian ethical narratives (Tasioulas, 2018). Thus, AI ethics requires that one addresses both conceptual and practical concerns.

In conclusion, the disparity between AI's actual capabilities and public perceptions with different conceptual and practical concerns underscores the need for balanced considerations that drive AI ethics attempts based on ethical frameworks. While AI's actual limitations must inform ethical guidelines, addressing public concerns and misconceptions is equally crucial to fostering responsible AI development and deployment (e.g., Bleher & Braun, 2023; Wong, 2021). Thus, in order to establish ethical frameworks, such hybrid approaches must ensure that AI ethics frameworks are both practically grounded and socially responsive.

4.3 Perceptions of HMA and AMA that may disrupt traditional notions of moral agency

The following section dives deeper into perceptions of morality in the interaction between humans and AI that vary across different research streams, thereby corresponding to our third RQ: *"In what ways might AI transformation challenge traditional notions of moral agency and responsibility?"* Table 12 in Appendix XIII presents a concise summary of the key findings, including percentages.

To what extent can AI systems embody or exhibit a virtuous moral character, and what implications does this have for our understanding of responsibilities?

The question of whether AI systems embody or exhibit a virtuous moral character differs in the literature, such as the terms used. The literature often refers to the Moor (2006) definition of explicit and implicit ethical agents in the context of machine ethics, which contains a long and controversial debate on the development of ethical machines that follow moral decision-making principles (Martinho et al., 2021). Thus, the debate on whether machines can behave ethically and drive ethics to some extent is an old one, often related to the idea of "singularity" (Boyles, 2018, p. 182), i.e., the idea of machines being built that are "powerful enough to replicate human moral agency" (Manna & Nath, 2021, p. 145).

Some of the literature analyzed suggests that AI technologies can only be explicit ethical agents, operating based on programmed principles and having limited moral agency (e.g., Farisco et al., 2020; Kim et al., 2018; Koebis et al., 2021; Taeihagh, 2021). While it is conceivable that strong AI could exhibit characteristics similar to full ethical agency, the absence of AI's free will is stated to imply a fundamental disparity between AMA and HMA (Van De Voort et al., 2015). Some authors emphasized that HMA is distinguished by its emotionality, a characteristic that sets it apart from AMA (e.g., Giroux et al., 2022; H. Roberts et al., 2023), and its consciousness (e.g., Lloyd, 1985; J. S. Roberts & Montoya, 2022). As Dyrkolbotn (2018, p. 79) framed it: "The agent can still be highly autonomous and highly ethical, but the underlying causes of its ethical decisions must be traced back to human agency." Another paper argues that "Machines do not have an inherent conscience or morality as humans do" (J. S. Roberts & Montoya, 2022, p. 19). This is supported by Saariluoma and Leikas (2020), who outline that AI can have ethical rules but no ethical norms since one cannot derive values from facts. They suggest that "it is easy to provide ethical rules for intelligent machines [...], but there are no principled obstacles to programming

ethical norms into machines and making them follow these norms” (Saariluoma & Leikas, 2020, p. 597). This can further lead to questions such as whether we can program AI rules that incorporate certain moral values and ascribe the developer’s responsibility despite potential misuse by the user.

By contrast, a minority suggests that AI technologies can exhibit implicit ethical agency, particularly in unsupervised learning systems capable of ethical reasoning based on human data input (Cook, 2023; Saariluoma & Leikas, 2020). One paper argues that “If human ‘interrogators’ cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent” (Allen et al., 2000, p. 354). Thus, even if the authors perceive AI as not having morality in a human sense since HMA can be perceived as unique because it is based on emotionality (Giroux et al., 2022; J. S. Roberts & Montoya, 2022), some literature suggests that considerations on ethics need to see AIs as active participants and ethical decision-makers to guarantee ethical interactions. For instance, Hauer (2022, p. 1) argues that “if a task requires some form of moral authority when it is performed by humans, its full automation, transferring the same task to autonomous machines, platforms, and AI algorithms, necessarily implies the transfer of moral competence.” Similarly, Millar (2016, p. 791) noted, “By treating the robot as an active participant in a human-robot relationship [...] we can characterize robots that make decisions for users in hard ethical cases as moral proxies.” Understanding AI technologies as moral proxies can foster ethical considerations in human-AI interaction because “the ethical debate surrounding moral proxies is well established, especially in the healthcare context.” (Millar, 2016, p. 791) In addition, research shows that, to some extent, humans attribute moral agency to AI, for instance, due to its human-like behavior and its ability to mimic human friendship (Li, 2021; Parlangeli et al., 2020; Willems et al., 2022) and since AI is reflecting their values (Seymour, 2018), leading to a perception of equivalence between HMA and AMA, which may inadvertently limit their moral agency (Shank et al., 2019; Tasioulas, 2018). Therefore, realistic perceptions regarding AMA are needed.

Conversely, a larger proportion of authors argue that AI technologies cannot be implicit ethical agents because they lack free will, thus affirming the enduring responsibility of humans (e.g., Swanepoel, 2021). Ong (2021, p. 3), for instance, suggests that “current AI do not take [context] into account, and so may even be fundamentally limited in this endeavor of reading internal states solely from external behavior.” Similarly, Niederman and Baker (2023, p. 10) wrote: “We consider the escalation of the distancing of human agency from the effects of mechanical actions. There is a difference between the direct effects of purposeful human activity and indirect effects of machinery that responds to unintended actions or neglectful inattention rather than explicit, purposeful agency.” One paper even argues against the presence of AMA to safeguard human well-being (Bankins & Formosa, 2023). Nevertheless, other papers suggest that the transformation of a technical entity into an artificial, psychological, moral agent would necessitate the integration of various attributes, such as “intuition and empathy, intelligence, autonomy and desires, sentience, and behavioral equivalence, noting problems that arise in each case (Shevlin, 2021, p. 416). Furthermore, the literature highlights that blurred responsibilities remain a challenge, stating that “where duty falls when algorithms have varied degrees of agency to act is another problem that arises with the application of AI” (Pansoni et al., 2023, p. 1152). Autonomous weapon systems (e.g., Gómez De Ágreda, 2020; Kim et al., 2018), for instance, are attributed to another “degree of agency” compared to virtual teaching assistants (e.g., Eicher et al., 2018). The literature emphasizes the need to “establish an AI accountability mechanism to clarify the responsibilities of developers, users, beneficiaries, etc.” (H. Roberts et al., 2023, p. 78) since “in the real life we cannot at least not yet hold artificial intelligence and its algorithms accountable for the decisions it

makes” (Nevanperä et al., 2021, p. 2333). According to the authors’ perspectives, establishing robust AI accountability mechanisms as legally defined responsibilities (e.g., Colmenarejo et al., 2022; Wu et al., 2020) is crucial for addressing the call. Rismani and Moon (2023) added the relevance of fostering one’s competencies to overtake responsibilities besides their role, making them responsible.

Who strives better for moral implications, and what implications does this have for our understanding of responsibility?

Despite acknowledging the potential moral superiority of AI, some of the literature asserted that humans remain the only full ethical agents, even if AI may surpass some individual humans in morality (e.g., Serafimova, 2020; J. Zhang & Zhang, 2023). Humans are assumed to remain the only full ethical agents “because an agent is capable of norm-adherence in virtue of norm-endorsement, [and] AI systems are not capable of this” (Swanepoel, 2021, p. 162). Consequently, some authors suggest that HMA needs to control AI (e.g., Ong, 2021) since it “will remain in the foreseeable future dependent on human ethical scrutiny as well as on the moral and legal responsibility of humans” (Dodig Crnkovic & Çürüklü, 2012, p. 63). Even though AI can imitate responsible behavior, it lacks genuine moral agency, underlining that moral responsibility is inherently a human trait (Smith & Vickers, 2021). Moreover, Serafimova (2020, p. 9) suggests that even though AMA may “become more moral than humans once,” this “moral perfection” is, at a minimum, based on moral updates through humans. This is in line with Palmer & Schwan’s (2024, p. 1) suggestion of “more process, less principles,” highlighting the need to rather consider perceptions of morality throughout AI’s lifecycle, forming responsibilities between various stakeholders instead of discussing responsibilities for single principles. Also, Stahl et al. (2023) argue that ethics and human rights in AI are very complex and lack agreement on responsibilities, wherefore more than principle-based ethics is required.

However, some of the papers suggest that as AI progresses, AMA becomes more akin to HMA (Martinho et al., 2021). Tóth et al. (2022), for instance, differentiate between weak AI ethical agents that lack the capacity to recognize unethical decisions, moderate AI ethical agents that may imitate (un)ethical decisions, and strong ethical agents that may prioritize other objectives over the avoidance of harm, thereby potentially disregarding ethical considerations. Although we recognize that AI technologies and ethics cannot be easily categorized into rigid classifications of weak, moderate, and strong ethical agents – especially given the diverse perspectives on what constitutes ethics and AI – the distinction nevertheless highlights the possibility that different types of AI technologies may embody varying forms of ethical agency (Tóth et al., 2022). There is, however, no consensus among moral theorists (Allen et al., 2000) who interpret and use the terms “ethical agency/agents” and “moral agency/agents” in a variety of ways.

Another stream of research also emphasizes the importance of aligning AMA with the moral values of HMA to ensure ethical human-AI interaction (e.g., Minkkinen & Mäntymäki, 2023; Vică et al., 2021; Wright, 2020). According to Telkamp and Anderson (2022), successfully implementing this would require enhancing our understanding of human values, which involves navigating conflicting expectations regarding moral values. This requires the integration of diverse human agencies. Moreover, applying moral foundation theory to nonhuman technologies with some inherent agency serves to deepen our understanding of moral values by differentiating between ethical and unethical mechanisms, facilitating their transferability to AI systems. Van de Poel (2020) emphasizes the importance of monitoring and redesigning AI systems to ensure constant alignment with human

values. In line with that, other papers advocate for the primacy of HMA in regulating and controlling AI, suggesting that AMA should not replace HMA (Knox, 2023; Neubert & Montañez, 2020). One paper argues that “Despite many discussions about technical morality, the moral subject is still assigned to human designers or users” (Sun & Ye, 2023, p. 9). Furthermore, Hauer (2020) addressed philosophical anti-dualism, meaning that the moral capabilities of AI technologies are limited to the ethical understanding of their developers. Thus, while AI technologies and robots can be programmed to do morally right things, behaving ethically overall is more than adhering to a specific code (Génova et al., 2023; Obi & Gray, 2023). Besides, there is ambivalence about AI with moral capabilities that can be “useful to humans [but] also have the capacity to do things that are harmful to humans” (Allen et al., 2000, p. 251). Therefore, Arnold and Schuetz (2018) argue that human control afterwards is not enough as interventions to avoid harmful consequences must occur earlier when the system is learning based on human behavior.

By contrast, other literature argues that HMA is somehow limited. Some papers argue that HMA may also have certain limits since “humans may be poor moral arbiters, not suitably wired to solve difficult ethical dilemmas,” following the virtue ethics view. Seo and Thorson (2022) suggest that human ethical and moral reasoning is fundamentally computational and does not involve some special human-only force or quality, which suggests a more deontological view. Other literature reveals concerns regarding the opacity of AI ethical principles and logic: “If the AI utilizes a black box algorithm [it is hardly] feasible to assess the factors which the AI utilizes” (Pruski, 2023, p. 11). Consequently, opacity regarding an AI’s ethical principles (explicit agents) or ethical logic (implicit agents) can lead to moral uncertainties among humans and potentially limit HMA (e.g., Niederman & Baker, 2023; Pruski, 2023). While “AI systems ‘learn’ from data and can self-evolve over time due to exposure to new data, without explicitly being programmed by a human” (Munoko et al., 2020, p. 2), “AI might be limited when it comes to complex judgments that require professional skepticism.” (Munoko et al., 2020, p. 4). As a consequence, HMA may be limited when AI’s architecture is invisible (Susser, 2019). Furthermore, “it may be unreasonable to expect designers to fully anticipate the behaviors of code that have the potential to suppress freedom” (McDonald & Pan, 2020, p. 147:3). In particular, “in complex systems, agents interact with each other and with other systems in unexpected ways, making their response to change unpredictable and non-linear” (Morley et al., 2021, p. 253). Thus, responsibility gaps can further arise when AI’s opacity limits HMA.

In sum, we suggest that instead of comparing AMA and HMA, theorizing about AI should be focused on how to combine the benefits of both. Researchers must consider the specific type of AI technologies and their capabilities since this determines AI’s autonomy and how individuals interpret its moral agency.

How do the interdependencies between HMA and AMA necessitate a framework of shared responsibility among different stakeholders?

Some of the literature argues that both AMA and HMA are necessary for ethical human-AI interaction, as neither alone provides a sufficient benchmark (e.g., McDonald & Pan, 2020; Siau & Wang, 2020). The AI ethics should be addressed within the design and not as an afterthought because AI should help humans by facilitating their work (Horváth, 2022). Some papers even suggest that AI can contribute to strengthening HMA by adhering to ethical principles, thereby influencing human moral perceptions (e.g., Schlimbach & Khosrawi-Rad, 2022; Willems et al., 2022). Since AMA can be understood as being based on norm adherence to programmed principles (Swanepoel, 2021; J. Zhang & Zhang, 2023), AI technology can be an effective way to train people on moral

issues if it makes them reflectively decide on their own, rather than being directly influenced (Lara, 2021). Mayer et al. (2021, p. 8) write that “AI ethics guidelines can serve as decision-making support for employees in critical cases. Employees then do not have to be concerned about whether decisions are ethically justifiable from their own point of view but can rely on the guidelines in making and justifying their decisions.” Interacting with AI may increase the awareness of ethical issues, thereby influencing HMA – although the AI itself may not have its own moral agency (Shih et al., 2021). Furthermore, AMA can be used to strengthen HMA since the system architecture that proposes AMA can help solve ethical dilemmas through collaboration (Córdova & Vicari, 2022).

However, a certain consensus emerged on the role of HMA as the ethical baseline for AI design and governance, advocating for shared responsibility among various human stakeholders (e.g., Khan et al., 2022; Vakkuri et al., 2019, 2021; Volkman & Gabriels, 2023). The suggested shared agency between AI and humans is argued to be due to the complex chain of interactions. Moral responsibility, therefore, can be attributed to companies, developers, or the AI system (Sullivan & Fosso Wamba, 2022). However, some of the literature suggests that “humans remain the duty bearers” (J. Zhang & Zhang, 2023, p. 1) and that risk should be distributed “fairly among users” (Yu et al., 2018, p. 5539). Moreover, regulations by governmental institutions in different countries provide a form of practical guidance on ethical AI principles (Tidjon & Khomh, 2023) and serve as valuable platforms for sharing learning and experiences across various governance initiatives (Auld et al., 2022). Yet, research highlights that “there is an underlying tension between individual agency and bureaucratic control with actual responsibility lying somewhere in between” (Raji et al., 2021, p. 517). In addition, our analysis found that other researchers believe that the development of AMA depends on HMA, highlighting the interdependence between HMA and AMA (e.g., Bryson, 2018; Simpson, 1992). Terzis (2020, p. 221), for instance, suggests the attribution of agency to AI constitutes a form of “fictional personification,” fostering a misconception that subjective values can be implemented in AI agents while, in reality, the “ethics of AI” reflects the ethics of its creators. It should be noted that while data and humans influence AI morality, humans can both foster morality in AI and manipulate it (Schneider et al., 2023). Thus, it is sometimes about the way in which AI is used that influences moral conduct. In response, the literature suggests strengthening HMA.

In order to manage AMA responsibly, some papers emphasize the need to strengthen HMA and clarify responsibilities in human-AI interaction (e.g., Buhmann & Fieseler, 2023; Burema et al., 2023). This seems important given that a robot’s ethical design can be transformed by humans, as “both ethical and unethical behavior require the same cognitive machinery” (Vanderelst & Winfield, 2018, p. 320). Thus, the human moral agency can be shaped through the training of moral behavior. This should encompass enhancing moral sensitivity abilities to recognize morally significant facets, developing moral judgment abilities to select the most appropriate course of action, fostering moral motivation to adhere to moral obligations, and cultivating moral action abilities to articulate, advocate, and justify moral decisions (Vanhée & Borit, 2022). Other suggestions to influence HMA include the implementation of a carrot-and-stick mechanism (Wang et al., 2020). While the latter can enforce norm compliance, there is also a need for further investigation into how differences in social norms and regulatory environments across regions affect the feasibility and acceptance of universal ethical standards. Ethical AI debate clubs could be one approach to foster cross-functional thinking about moral concerns and ethical dilemmas that arise with AI and the establishment of a Chief Responsible AI Officer in organizations (Wang et al., 2020). Additionally, exchanging ideas about how to develop the most unethical AI agent can also help to emphasize ethical considerations besides principles of AI ethics (Weiss et al., 2021).

In conclusion, the literature suggests that there are certain interdependencies between HMA and AMA that require a specific splitting of responsibilities (e.g., Khan et al., 2022; Volkman & Gabriels, 2023). Given the complex interactions between AI and humans, this shared responsibility is perceived as crucial, but it ultimately affirms that humans must remain the primary bearers of moral duties. In order to manage AMA responsibly, some literature emphasizes the need to strengthen HMA through training, ethical debates on unethical AI, and organizational roles, such as a Chief Responsible AI Officer (e.g., Henderson, 2019; Wang et al., 2020; Weiss et al., 2021).

Discussion

In this section, we outline the theoretical implications of our work and provide three avenues for future research (Section 5.1). We also discuss the practical implications and provide recommendations for various stakeholder groups to enhance AI ethics within different AI application domains and disciplines (Section 5.2). Finally, we end with limitations and suggestions for future work (Section 5.3).

5.1 Theoretical Implications

Our review results provide theoretical contributions, each highlighting avenues for future research. Below, we focus on the main contradictions highlighted by the literature reviewed that appear across different views on ethics and synthesize them to provide key implications for future work on AI ethics.

How fragmented ethical views in AI research can benefit from empirical studies grounded in descriptive ethics

As highlighted in our review results, particularly in Section 4.1, there is a substantial variation in the application of AI ethics frameworks, conceptualizations, and principles, with many studies relying on specific ethical viewpoints. This has led to a geographically and academically bounded dialogue on ethics in human-AI interaction (e.g., Colmenarejo et al., 2022; Stahl et al., 2022; Wu et al., 2020). Such a limited approach fails to encompass the diverse moral communities worldwide, underscoring a need for a broader interdisciplinary and international dialogue to establish geographically aligned ethical standards and conceptualizations (e.g., Rismeni & Moon, 2023; Telkamp & Anderson, 2022). Since “[m]orality is a relative concept, which changes significantly with the environment and time” (Sun & Ye, 2023, p. 1), there is a need to consider sense-making across several domains – AI design, regulation, deployment, education, and beyond – to develop a more nuanced understanding of AI ethics. Palmer and Schwan (2024) argue that this requires a deeper focus on the cognitive and behavioral processes behind human-AI interaction rather than an exclusive focus on specific deontological principles.

Empirical studies grounded in descriptive ethics can play a crucial role in enhancing our understanding of differences in AI ethics research (e.g., Tóth et al., 2022). Particularly, descriptive ethics can help to understand the complex and coevolutionary relationship between regulation (linked to deontology) and human behavior (linked to virtue ethics) (e.g., Erman & Furendal, 2022; H. Roberts et al., 2023). Regulation with given principles per duty can influence human behavior, prompting reflection and the adoption of corresponding maxims, and vice versa (e.g., Erman & Furendal, 2022). Thereby, regulation should not solely aim at a consensus between a human moral maxim and ethical rules. As Jones et al. (2005, p. 176) asserted, “[t]he moment that ethics stop questioning these rules, it stops being ethics.” Hence, an excessive focus on alignment might fail to capture the diversity of

moral viewpoints that exist among moral communities (Constantinescu et al., 2021; Meijer et al., 2023). Two avenues for future empirical investigations grounded in descriptive ethics are:

1. **Examining correlations between academic domains or nations and their positions on AI:** Empirical studies, such as large, global online surveys, should explore whether certain academic disciplines have consistent positions regarding AI, potentially revealing patterns of ethical interpretation that could inform a more nuanced understanding of global AI ethics.
2. **Investigating how cultural contexts shape ethical perceptions of AI among research disciplines and across nations:** Empirical experiments could explore how cultural contexts and research disciplines influence ethical perceptions of AI technologies and mechanisms to minimize risks, providing insight into how different societies and research disciplines prioritize values such as privacy, autonomy, and fairness in similar AI application scenarios. An intriguing line of inquiry would be to explore the extent to which risk mitigation strategies of several nations – such as the implementation of technical safeguards and other protective measures – are themselves perceived as ethical concerns across nations or whether they are predominantly viewed as necessary steps to reduce risks. An additional avenue of research could investigate how various research disciplines develop ethical perceptions of AI by constructing theoretical foundations grounded in referencing highly cited publications within each field.

How the establishment of hybrid approaches for ethical AI frameworks can significantly benefit from consensus-driven ethical deliberations

Our findings, particularly in Section 4.2, indicate that the development of ethical frameworks guiding human-AI interaction is fraught with challenges due to the presence of various, often conflicting, approaches. Hybrid approaches have been proposed to address the complexities and contradictions inherent in ethical frameworks, necessitating a structured discourse to reconcile these differences. Habermas' discourse ethics, as suggested by Vică et al. (2021), offers a valuable tool for legitimizing such frameworks through critical discussions that incorporate diverse ethical perspectives within society. This approach emphasizes the importance of inclusive and rational dialogue in reaching a consensus on ethical matters, ensuring that the resulting frameworks are both justifiable and reflective of the collective moral will (Habermas, 1984, 1993). The relevance of discourse ethics is further underscored by several other scholars in the literature we reviewed, including Niedermann and Baker (2023), Morley et al. (2021), and Westerstrand (2023). These authors highlight the necessity of finding common legal, ethical, and moral ground across different cultures while simultaneously respecting cultural differences. However, achieving this is a complex task, as even well-established frameworks can be interpreted in various ways depending on cultural contexts.

Involving Habermas' discourse ethics in the development of hybrid AI frameworks not only promotes consensus building but also ensures that there is an open discourse to identify and establish ethical truths that remain open to diverse perspectives, thus preventing the dominance of any single ethical viewpoint (e.g., Morley et al., 2021). When considering the regulation of AI in healthcare, for example, discourse ethics would encourage the inclusion of perspectives from medical professionals, patients, technologists, and ethicists from various cultural backgrounds (e.g., Emdad et al., 2023; Palmer & Schwan, 2024; Wehrens et al., 2023). This inclusive approach ensures that ethical guidelines are culturally sensitive and ethically sound, leading to more robust AI

ethics frameworks. However, as Niederman and Baker (2023, p. 23) argue “Habermas’ processes do not guarantee that applying discourse ethics approaches will work in all cases. Future research may refine these maxims and their corollaries. Still, we suspect it will be more valuable to show how they can methodically be applied to real problems and, even more importantly, lead to prescriptions of behavior that will prevent and ameliorate the harms that the use of AI systems may create.” Hence, two avenues for future work using Habermas’ discourse ethics as a methodological approach are:

1. **Exploring cross-cultural consensus building in AI Ethics:** Future research could investigate how Habermas’ discourse ethics can be applied to facilitate cross-cultural consensus building when establishing AI ethics frameworks. This includes assessing the effectiveness of discourse ethics methods, such as Delphi studies with focus group interviews that bring together stakeholders from diverse cultural backgrounds to participate in structured dialogues, shaping inclusive and widely accepted AI policies.
2. **Investigating interdependencies between legal, moral, and ethical frameworks through discourse ethics:** Another key area of research involves exploring the interdependencies between legal, moral, and ethical frameworks in the context of AI through discourse ethics. This could involve mapping studies analyzing existing frameworks and structured dialogues, such as through crowd-working exercises, to understand how frameworks influence one another and shape AI design, deployment, and utilization.

How the alignment of different perceptions of HMA and AMA benefit from the process of reconciling contradictory perspectives

Our findings, particularly Section 4.3, suggest a shift in the academic discourse from merely comparing AMA with HMA toward a more integrative approach that seeks to reconcile their differences to reveal new possibilities for reducing responsibility gaps in human-AI interaction. As AI systems become increasingly capable, ensuring that these systems are trained on data reflecting current societal compositions is critical (Pruski, 2023). However, the concept of agency and responsibility in the context of AI remains elusive for many, as AI’s moral agency is deeply intertwined with societal values and inherently shaped by human design (Dyrkolbotn et al., 2018; Niederman & Baker, 2023). In the context of increasingly autonomous, self-learning, and human-like AI technologies, the complex relationship between HMA and AMA leads to moral uncertainties (Li, 2021), where the boundaries of responsibility become indistinct. This interplay between HMA and AMA presents both interdependencies and tensions, as humans often perceive AIs as moral proxies while simultaneously experiencing uncertainty due to the opaque nature of AI decision-making (Coeckelbergh, 2020; Niederman & Baker, 2023; Pruski, 2023; Shevlin, 2021). Thus, to pick up Floridi and Sander’s (2004) arguments, AI extends traditional boundaries of moral involvement by being both moral patients and moral agents. Nevertheless, AMA can also augment HMA by demonstrating superior moral reasoning in specific tasks (Serafimova, 2020; J. Zhang & Zhang, 2023). The evolving and sometimes contradictory perceptions of HMA and AMA create moral ambiguities and blurred lines of responsibility in human-AI interactions. The challenge lies in understanding and navigating the interdependencies and tensions between HMA and AMA, particularly as AI systems gain capabilities that may surpass human moral reasoning in certain contexts.

In order to address these challenges, we propose a shift in the academic discourse from a comparative analysis of HMA and AMA to a more integrative approach that seeks to reconcile their differences. This can help to mitigate moral uncertainties in human-AI interactions. Additionally, this might help to clarify the often similarly used concepts of ethical agents, moral agents, and moral patients (e.g., Allen et al., 2000), particularly given the ongoing advancements in AI technology (Dwivedi et al., 2023). Two avenues for future work are:

1. **Longitudinal empirical investigations of the evolution of moral norms in human-AI interaction:** Future research should explore how advancements in AI technology impact the ongoing reconciliation of human and artificial moral agencies, involving perceptions of different stakeholders, such as developers, managers, and users. These studies should examine how stakeholders perceive the moral agency of AI when AI capabilities continue to develop and how these perceptions influence their interactions.
2. **Developing a conceptualization of interdependencies between AI and human agency:** Another avenue for future research is the development of a conceptualization of AI and human agency in general and their specific notions of moral agency (HMA and AMA), moral patiency, and ethical agency. This conceptualization could involve ontological differences between AI ethics researchers and integrating insights from philosophy, ethics, law, and computer science.

5.2 Practical Implications

Our findings highlight that literature on AI ethics spans a broad spectrum of application areas, each offering distinct recommendations for a range of stakeholders. Below, we outline key recommendations for different stakeholder groups (for further details, see Appendix XIV; Table 13).

Regulators should be aware of the prominence of deontological principles of AI governance at different levels

Most of the articles reviewed are closely related to the literature on AI governance, particularly in the context of the establishment of ethical standards, norms, principles, and laws for the design, deployment, and usage of AI across multiple layers (e.g., Almeida et al., 2020; Henderson, 2019; Tidjon & Khomh, 2023; Walke et al., 2023). AI governance manifests at different levels, such as technical, industrial, and societal ones (e.g., Shneiderman, 2020). Regulation at the technical level, for instance, may encompass AI design principles, while at the industrial level, it may involve the interdependent oversight of AI technologies. At the societal level, governance can take the form of laws, such as the EU AI Act (Hagendorff, 2020; Koniakou, 2023; Shneiderman, 2020). This complexity necessitates structured approaches that ensure the alignment of ethical principles across these different layers. Regulators and policymakers need to recognize that beyond established laws and norms on a societal level, often rooted in deontological ethics, different cultural contexts may embody distinct societal moral values linked to virtue ethics (Agbese et al., 2023; Seppälä et al., 2021). Similarly, conflicts may arise across industrial levels, where some sectors, such as medicine, are highly regulated while others, such as manufacturing, have fewer ethical sensitivities (Mäntymäki et al., 2023). Consequently, imposing uniform standards at the industrial level might stifle innovation in some organizations due to the pressure to meet legal requirements, even though these standards are crucial for others (e.g., Mayer et al., 2021). Thus, the predominance of ethical principles in AI governance can become counterproductive when not properly aligned with the diverse values present in various cultural contexts. Townsend et al. (2022) recommended bringing normative principles more into practice (e.g., by

sector-specific regulations) and addressing social, legal, and cultural considerations to promote virtues besides principles. This approach might help to establish AI ethics frameworks that establish a layered network of moral maxims and ethical principles tailored to distinct cases as a universal, one-size-fits-all set of AI principles is unlikely to accommodate the diverse cultural and moral contexts and would, at best, achieve a minimum ethical baseline for AI regulation (e.g., Eicher et al., 2018).

Leaders should integrate virtue-based approaches within their AI deployment strategies

The integration of AI ethics into business processes necessitates the development of AI strategies that align with both worker and AI values (e.g., Kieslich et al., 2022; Van De Poel, 2020) and provide guidelines that promote ethical behavior in human-AI interactions (e.g., Bankins & Formosa, 2023; Pruski, 2023). Our review emphasizes the importance of maintaining human control over AI, particularly in sensitive or high-risk contexts such as the military, where transparency in and explainability of algorithmic operations and human oversights are crucial (Ashrafian, 2015b; Gómez De Ágreda, 2020; Swanepoel, 2021; Walke et al., 2023; Weber, 2020). Additionally, aligning AI systems with societal values is vital to prevent harm and unintended consequences, with organizations bearing responsibility for negative outcomes resulting from AI deployment (Agbese et al., 2023; Buhmann & Fieseler, 2023; Rodgers et al., 2023; Taeihagh, 2021). It is essential to integrate virtue-based approaches within HRM practices, fostering responsibility-taking (Buhmann & Fieseler, 2023; Burema et al., 2023; Hagendorff, 2020; Rodgers et al., 2023). For instance, organizations can cultivate employee virtue ethics by establishing ethical AI debate clubs to foster cross-functional thinking about moral concerns (Wang et al., 2020). This approach not only helps in managing the relationship between perceived harm and blame judgments (Sullivan & Fosso Wamba, 2022) but also supports a morally sound and responsible work environment (e.g., Burema et al., 2023; Hagendorff, 2020; Sullivan & Fosso Wamba, 2022).

Educators should account for an interdisciplinary view of AI ethics in classrooms

Our review emphasizes the critical need to integrate ethics into AI curricula, highlighting its close connection to the concept of AI literacy (Schneider et al., 2023). Developing AI-related skills, knowledge, and competencies is essential for not only guiding educators in effectively teaching AI ethics (Raji et al., 2021) but also fostering students' ethical skills (Williams et al., 2021). As AI capabilities advance, educators require novel pedagogical approaches to AI ethics to ensure they are adequately “prepared and successfully engage[] their students in the topic” (Williams et al., 2021, p. 15678). The integration of ethics into AI education is particularly important, given the potential ethical pitfalls of AI systems and the need to enhance knowledge about the differences between HMA and AMA (Horváth, 2022; Lara, 2021; Pruski, 2023). Weiss et al. (2021, p. 331) outlined that instead of focusing on existing ethical principles, students might design the “most unethical chatbot imaginable” to derive ideas on what is a responsible design of AI. Thereby, domain-specific circumstances should be considered, such as using AI systems in a medical context requires considering the unique characteristics of the AI technologies that are used in healthcare (e.g., Katznelson & Gerke, 2021). Furthermore, the literature outlines the relevance of collaboration across various fields of education, as AI ethics is underrepresented in computer science and IS education (e.g., Williams et al., 2021). Moreover, when AI is further used to assist teaching, such as with pedagogical conversational agents (Schlimbach & Khosrawi-Rad, 2022) or virtual teaching assistants for frequently asked questions that may be used in classrooms (Eicher et al., 2018), such assistive AI technologies for education might struggle when it comes to unexpected situations. For instance, if an underaged student, for

instance, shares that she is pregnant and the AI assistant simply congratulates her (Eicher et al., 2018), it is crucial to involve well-educated human teachers or trainers to explain AI's outputs (Williams et al., 2021). AI ethics education must be reconsidered and given greater interdisciplinary recognition to meaningfully address these challenges (Raji et al., 2021).

Developers should link deontological principles with a human-centered perspective on AI design

The literature reviewed highlights that AI developers should prioritize a human-centered approach in the design of AI technologies, ensuring that human needs and values are central to their work (e.g., Shneiderman, 2020). This involves considering the ethical, economic, and societal implications of AI technologies from the outset, particularly in unsupervised learning systems capable of ethical reasoning based on human data input (Cook, 2023; Saariluoma & Leikas, 2020). Developers should incorporate established ethical frameworks to guide decision-making and anticipate potential ethical dilemmas (Fukuda-Parr & Gibbons, 2021; Simpson, 1992) and should use design methods that build upon AI ethics research to implement ethically aligned AI systems in different contexts (Vakkuri et al., 2021). It is essential to implement responsible AI design guidelines that include security and accuracy backups, thereby minimizing the risks of unintended consequences and effectively managing trade-offs between different design choices (Buruk et al., 2020). Hence, AI developers should emphasize “transparency, fairness, equality, explainability and avoiding bias or discrimination” throughout the AI lifecycle (Carneiro & Veloso, 2022, p. 1). Practical recommendations include the adoption of robust testing procedures that simulate real-world scenarios, continuous monitoring of AI systems in operation, and the inclusion of diverse teams in the development process to identify and mitigate potential biases (e.g., Arnold & Scheutz, 2018; Munoko et al., 2020). By embedding these principles into their work, AI developers can create technologies that are not only effective but also socially responsible and trustworthy (e.g., Shneiderman, 2020).

Researchers should investigate the interconnection between HMA and AMA

It seems to be important to determine the moral responsibilities of HMA and AMA based on the interdependencies that arise in a specific interaction between humans and AI rather than trying to define HMA and AMA characteristics separate from these interaction dynamics. Our review shows that a great amount of research addresses responsibility gaps that arise when, for instance, AI's mechanisms are opaque and unpredictable (e.g., Buhmann & Fieseler, 2023; Gahnberg, 2021), responsibilities are spread among several human actors involved in any way before or during the human-AI interaction, such as developers and managers, and clear commitments among these actors are absent. The debate on the moral responsibilities of HMA and AMA is especially crucial because consequences happen asynchronously and may be delayed. It is critical to note that AI failures may occur years later when AI developers are no longer around, making it difficult to assign blame (e.g., Constantinescu et al., 2021; Smith & Vickers, 2021); human users' awareness of and knowledge about accountability may not exist or, worse yet, may be ignored (Champagne & Tonkens, 2023; Coeckelbergh, 2020). Such dynamics in a specific interaction context influence how human users attribute responsibility to themselves or others (Coeckelbergh, 2020; Constantinescu et al., 2021). Therefore, researchers are encouraged to conceptualize possible interdependencies that impact perceptions of HMA and AMA and, thus, influence how moral responsibilities are attributed. Therefore, Zhang et al. (2021) recommended focusing on examining the factors that drive the heterogeneous constitution of AI ethics research by building international scholarly collaborations.

In conclusion, our review emphasizes the importance of nuanced and interdisciplinary approaches to AI ethics, resulting in tailored recommendations for various stakeholders. Advancing AI ethics requires collaborative efforts across disciplines to navigate the complex and multifaceted nature of ethical human-AI interaction.

5.3 Limitations and Future Research

This work has some limitations that also point to future research directions.

Firstly, while we aimed to provide a holistic and systematic overview of the academic dialogue on ethics within AI research, our findings are restricted to this previous knowledge of scholarly communities. Following the idea of Burton-Jones et al. for “Next-Generation Theorizing” (2021, p. 306), the findings of this work could be used in combination with insights from practitioners. Current debates on how to integrate ethics when designing, deploying, using, managing, regulating AI, and using AI in education would benefit from a practical perspective. Qualitative studies involving AI engineers and managers, for example, could explore the risks they identify and the measures they implement to establish safeguards or improvements. This would enhance how we establish theoretical frameworks and conceptualizations on AI ethics and ethical human-AI interaction.

Secondly, the objective of understanding how the use of ethical theories impacts perceptions of moral agency and responsibilities within AI research calls for an approach that favors breadth over depth. Even though we demonstrate relationships between normative ethical theories and interpretations of HMA compared to AMA, we focused on providing a holistic and overarching picture. Future research could zoom in on specific relationships that have emerged from our analysis with an empirical investigation coupled with experimental analysis, such as examining the interdependencies between the perceived emotionality of AI technologies and how users attribute moral agency to the AI because of its ability to imitate human-like behavior. Another possibility for future research would be to examine individuals’ cognitive processes when evaluating ethical aspects of AI technology creation, to determine whether certain ethical frameworks are referenced, and to identify any additional considerations that may surface.

Thirdly, 86.09 % of the articles reviewed have appeared in the last five years, which highlights that ethical considerations within AI research are booming, even though the dialogue is somehow spread across research disciplines and nations. Given that our review is restricted to a certain time range and we have not yet conducted backward and forward searches, future work might continue our work with systematic review updates that would capture new academic insights on ethics within AI research published after 2023. In addition, given that there are few articles from the early 2000s or earlier that address ethical issues in the context of AI research, future work could further compare our findings with the use of ethical theories in the machine ethics literature. Finally, involving further disciplines, such as political sciences, would enhance perspectives on how we perceive and understand ethical AI deployment, usage, and management.

Conclusion

With AI technologies demonstrating increased self-learning, (semi-)autonomous, and human-like capabilities, there is a profound transformation taking place in how humans interact with technologies. The evolving dynamics within human-AI interaction come with complex ethical challenges and blurred lines of responsibility, highlighting the need to increase our understanding of ethics within AI research by combining technical discourse with moral philosophy. Various ethical views impact our perceptions of morality within human-AI interaction.

Thus, this work examines those interdependencies within current AI-related research based on a systematic literature review.

Our findings reveal that AI ethics research is fragmented with clusters that predominantly focus on deontological principles, while others focus on virtue ethics, leading to different responsibility attribution behavior. To integrate both perspectives, deontological discourses that often reflect laws and regulations of specific communities, regions, and nations necessitate more empirical research on the interplay between regulations, societal values, and human behavior (linked to virtue ethics). Additionally, our findings outline challenges when establishing ethical frameworks due to conflicting approaches shaped by cultural and contextual factors. Fostering diversity of ethical perspectives and dialogue among different research disciplines is recommended to balance those challenges. Finally, our findings highlight different perceptions of humans compared to artificial moral agency. Instead of comparing perceptions of AMA and HMA, we suggest combining their strengths to achieve a more cohesive understanding of responsibility and morality within the AI context.

We hope this work provides a valuable starting point, encouraging researchers to reflect upon their use of ethical theories and their implications on agency perceptions and judgments of responsibility. As we navigate the labyrinth of AI research, guided by diverse ethical views, let us remember that our choices shape not only the development of technology but also our collective moral consciousness. The essence of our shared journey toward understanding accountability and moral and ethical agency in human-AI interaction lies at this intersection

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

All authors have read and agreed to the final version of the manuscript.

Availability of data and material

Not applicable

Competing interests

The authors do not have any conflicts or interests to declare.

Funding

No funding was received.

Authors' contributions

The authors confirm their contribution to the paper as follows: Conceptualization: T.H., A.H., K.S., and N.P.; methodology: T.H., A.H., K.S. and N.P.; investigation: T.H., A.H., K.S., and N.P.; writing – original draft preparation: T.H., A.H., K.S., and N.P.; writing – review and editing: T.H., A.H., K.S., N.P., O.P., and K.G.; supervision: O.P. and K.G.; project administration: T.H.

Acknowledgements

The authors would like to express their gratitude for guidance and support from the editorial team.

References

- Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006). Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis*. *Journal of Management Studies*, 43(5), 957–983. <https://doi.org/10.1111/j.1467-6486.2006.00625.x>
- Agbese, M., Mohanani, R., Khan, A., & Abrahamsson, P. (2023). Implementing AI Ethics: Making Sense of the Ethical Requirements. *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 62–71. <https://doi.org/10.1145/3593434.3593453>
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251–261. <https://doi.org/10.1080/09528130050111428>
- Almeida, P. G. R. de, Denner dos Santos, C., & Farias, J. S. (2020). Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance. *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Proceedings of the 53rd Hawaii International Conference on System Sciences.
- Anderson, M., & Anderson. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4).
- Arnold, T., & Scheutz, M. (2018). The “big red button” is too late: An alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, 20(1), 59–69. <https://doi.org/10.1007/s10676-018-9447-7>
- Ashrafian, H. (2015a). AIonAI: A Humanitarian Law of Artificial Intelligence and Robotics. *Science and Engineering Ethics*, 21(1), Article 1. <https://doi.org/10.1007/s11948-013-9513-9>
- Ashrafian, H. (2015b). AIonAI: A Humanitarian Law of Artificial Intelligence and Robotics. *Science and Engineering Ethics*, 21(1), 29–40. <https://doi.org/10.1007/s11948-013-9513-9>
- Ashrafian, H. (2015c). Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights. *Science and Engineering Ethics*, 21(2), Article 2. <https://doi.org/10.1007/s11948-014-9541-0>
- Auld, G., Casovan, A., Clarke, A., & Faveri, B. (2022). Governing AI through ethical standards: Learning from the experiences of other private governance initiatives. *Journal of European Public Policy*, 29(11), 1822–1844. <https://doi.org/10.1080/13501763.2022.2099449>
- Badea, C., & Artus, G. (2022). Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents. In M. Bramer & F. Stahl (Eds.), *Artificial*

- Intelligence XXXIX* (Vol. 13652, pp. 124–137). Springer International Publishing.
https://doi.org/10.1007/978-3-031-21441-7_9
- Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45(1), 315–341.
<https://doi.org/10.25300/MISQ/2021/15882>
- Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, 185(4), 725–740. <https://doi.org/10.1007/s10551-023-05339-7>
- Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, 28(3), 801–831. <https://doi.org/10.1086/345321>
- Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, 30(2), 195–218. <https://doi.org/10.1007/s11023-020-09525-8>
- Bentham, J. (1988). *The principles of morals and legislation*. Batoche Books Kitchener.
- Berber, A., & Srećković, S. (2024). When something goes wrong: Who is responsible for errors in ML decision-making? *AI & SOCIETY*, 39(4), 1891–1903. <https://doi.org/10.1007/s00146-023-01640-1>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). MANAGING ARTIFICIAL INTELLIGENCE. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Bilal, A., Wingreen, S., Sharma, R., & Jahanbin, P. (2021). Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics. *Proceedings of the Australasian Conference on Information Systems (ACIS)*.
- Bleher, H., & Braun, M. (2023). Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice. *Science and Engineering Ethics*, 29(3), 21.
<https://doi.org/10.1007/s11948-023-00443-3>
- Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines*, 27(4), 591–608. <https://doi.org/10.1007/s11023-017-9448-z>
- Boyles, R. J. (2018). A Case for Machine Ethics in Modeling Human-Level Intelligent Agents. *Kritike: An Online Journal of Philosophy*, 12(1), 182–200. <https://doi.org/10.25138/12.1.a9>
- Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. *Sustainability*, 13(4), 1974. <https://doi.org/10.3390/su13041974>

- Bryson, J. J. (2018). Patience is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26. <https://doi.org/10.1007/s10676-018-9448-6>
- Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. *Business Ethics Quarterly*, 33(1), 146–179. <https://doi.org/10.1017/beq.2021.42>
- Burema, D., Debowski-Weimann, N., von Janowski, A., Grabowski, J., Maftai, M., Jacobs, M., van der Smagt, P., & Benbouzid, D. (2023). A Sector-Based Approach to AI Ethics: Understanding Ethical Issues of AI-Related Incidents Within Their Sectoral Context. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 22(3), 705–714. <https://doi.org/10.1145/3600211.3604680>
- Burton-Jones, A., Butler, B. S., Scott, S. V., & Xu, S. X. (2021). NEXT-GENERATION INFORMATION SYSTEMS THEORIZING: A CALL TO ACTION. *MIS Quarterly*, 45(1), 301–314.
- Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, 23(3), 387–399. <https://doi.org/10.1007/s11019-020-09948-1>
- Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 155–205. <https://doi.org/10.1007/BF02019280>
- Carneiro, D., & Veloso, P. (2022). Ethics, Transparency, Fairness and the Responsibility of Artificial Intelligence. *New Trends in Disruptive Technologies*.
- Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The Sociomateriality of Information Systems. *MIS Quarterly*, 38(3), 809–830.
- Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26(2), 501–532. <https://doi.org/10.1007/s11948-019-00151-x>
- Champagne, M., & Tonkens, R. (2023). A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot harm. *Science and Engineering Ethics*, 29(4), 27. <https://doi.org/10.1007/s11948-023-00449-x>

- Chaput, R., Duval, J., Boissier, O., Guillermin, M., & Hassas, S. (2021). A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 13–23. <https://doi.org/10.1145/3461702.3462515>
- Cheruvalath, R. (2023). Artificial Intelligent Systems and Ethical Agency. *Journal of Human Values*, 29(1), 33–47. <https://doi.org/10.1177/09716858221119546>
- Churchill, L. R. (1999). Are we professionals? A critical look at the social role of bioethicists. *Daedalus*, 128, 253–274.
- Coeckelbergh, M. (2009). Distributive Justice and Co-Operation in a World of Humans and Non-Humans: A Contractarian Argument for Drawing Non-Humans into the Sphere of Justice. *Res Publica*, 15(1), 67–84. <https://doi.org/10.1007/s11158-009-9080-8>
- Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Colmenarejo, A. B., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., & Kinder-Kurlanda, K. (2022). Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 107–118. <https://doi.org/10.1145/3514094.3534158>
- Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). <https://doi.org/10.1609/aaai.v31i1.11140>
- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*, 23(4), 803–814. <https://doi.org/10.1007/s10676-021-09616-9>
- Cook, T. (2023). Robust Artificial Moral Agents and Metanormativity. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 162–169. <https://doi.org/10.1145/3600211.3604703>

- Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022). Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 864–876. <https://doi.org/10.1145/3531146.3533150>
- Cooren, F. (2004). Textual Agency: How Texts Do Things in Organizational Settings. *Organization*, 11(3), 373–393. <https://doi.org/10.1177/1350508404041998>
- Córdova, P. R., & Vicari, R. M. (2022). Practical Ethical Issues for Artificial Intelligence in Education. In *Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal, August 31–September 2, 2022, Revised Selected Papers* (Vol. 1720, pp. 437–453). Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-22918-3>
- Corvite, S., Roemmich, K., Rosenberg, T. I., & Andalibi, N. (2023). Data Subjects’ Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–38. <https://doi.org/10.1145/3579600>
- Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2019). Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. *Applied Artificial Intelligence*, 33(3), 267–293. <https://doi.org/10.1080/08839514.2018.1560124>
- Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial Intelligence (AI) and Information Systems: Perspectives to Responsible AI. *Information Systems Frontiers*, 25(1), 1–7. <https://doi.org/10.1007/s10796-022-10365-3>
- Dodig Crnkovic, G., & Çürüklü, B. (2012). Robots: Ethical by design. *Ethics and Information Technology*, 14(1), 61–71. <https://doi.org/10.1007/s10676-011-9278-2>
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette,

- Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Dyrkolbotn, S., Pedersen, T., & Slavkovik, M. (2018). On the Distinction between Implicit and Explicit Ethical Agency. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 74–80. <https://doi.org/10.1145/3278721.3278769>
- Eicher, B., Polepeddi, L., & Goel, A. (2018). Jill Watson Doesn’t Care if You’re Pregnant: Grounding AI Ethics in Empirical Studies. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 88–94. <https://doi.org/10.1145/3278721.3278760>
- Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1(1), 73–80. <https://doi.org/10.1007/s43681-020-00011-6>
- Emdad, F. B., Ho, S. M., Ravuri, B., & Hussain, S. (2023). Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence. *Twenty-Ninth Americas Conference on Information Systems*.
- Erman, E., & Furendal, M. (2022). The Global Governance of Artificial Intelligence: Some Normative Concerns. *Moral Philosophy and Politics*, 9(2), 267–291. <https://doi.org/10.1515/mopp-2020-0046>
- Estrada, D. (2018). Value Alignment, Fair Play, and the Rights of Service Robots. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 102–107. <https://doi.org/10.1145/3278721.3278730>
- Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, 21(4), 403–418. <https://doi.org/10.1007/s10892-017-9252-2>
- Fabiano, N. (2019). Robotics, intelligent systems, ethics and data protection: The new challenge. *Proceedings of the 2nd International Conference on Applications of Intelligent Systems*, 1–5. <https://doi.org/10.1145/3309772.3309787>
- Farisco, M., Evers, K., & Salles, A. (2020). Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence. *Science and Engineering Ethics*, 26(5), 2413–2425. <https://doi.org/10.1007/s11948-020-00238-w>
- Faulkner, P., & Runde, J. (2010). The Social, the Material, and the Ontology of Non-Material Technological Objects. *European Group for Organizational Studies (EGOS) Colloquium, Gothenburg*, 985, 4–8.

- Figueras, C., Verhagen, H., & Pargman, T. C. (2022). Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations. *Scandinavian Journal of Management*, 34(2).
- Flathmann, C., Schelble, B. G., Zhang, R., & McNeese, N. J. (2021). Modeling and Guiding the Creation of Ethical Human-AI Teams. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 469–479. <https://doi.org/10.1145/3461702.3462573>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Forsyth, S., Dalton, B., Foster, E. H., Walsh, B., Smilack, J., & Yeh, T. (2021). Imagine a More Ethical AI: Using Stories to Develop Teens’ Awareness and Understanding of Artificial Intelligence and its Societal Impacts. *2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, 1–2. <https://doi.org/10.1109/RESPECT51740.2021.9620549>
- Fosso Wamba, S., & Queiroz, M. M. (2023). Responsible Artificial Intelligence as a Secret Ingredient for Digital Health: Bibliometric Analysis, Insights, and Research Directions. *Information Systems Frontiers*, 25(6), 2123–2138. <https://doi.org/10.1007/s10796-021-10142-8>
- Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on ‘Ethical AI’: Human Rights Critique of Stakeholder Guidelines. *Global Policy*, 12(S6), 32–44. <https://doi.org/10.1111/1758-5899.12965>
- Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2), 218–231. https://doi.org/10.1162/daed_a_01911
- Gahnberg, C. (2021). What rules? Framing the governance of artificial agency. *Policy and Society*, 40(2), 194–210. <https://doi.org/10.1080/14494035.2021.1929729>
- Gaus, G. F. (2001). What is Deontology? Part One: Orthodox View. *The Journal of Value Inquiry*, 35, Article 35. <https://doi.org/10.1145/3594536.3595156>
- Génova, G., Moreno, V., & González, M. R. (2023). Machine Ethics: Do Androids Dream of Being Good People? *Science and Engineering Ethics*, 29(2), 10. <https://doi.org/10.1007/s11948-023-00433-5>

- Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, 36(1), 2009222. <https://doi.org/10.1080/08839514.2021.2009222>
- Gert, B. (2005). *Morality: Its Nature and Justification* (1st ed.). Oxford University PressNew York. <https://doi.org/10.1093/0195176898.001.0001>
- Ghatar, P. A., Pappas, I., & Vassilakopoulou, P. (2023). Practices for Responsible AI: Findings from Interviews with Experts. *Twenty-Ninth Americas Conference on Information Systems*. Twenty-ninth Americas Conference on Information Systems.
- Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>
- Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI. *Journal of Business Ethics*, 178(4), 1027–1041. <https://doi.org/10.1007/s10551-022-05056-7>
- Glänzel, W., & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. *Scientometrics*, 59(3), 281–310. <https://doi.org/10.1023/B:SCIE.0000018535.99885.e9>
- Gogoshin, D. L. (2021). Robot Responsibility and Moral Community. *Frontiers in Robotics and AI*, 8, 768092. <https://doi.org/10.3389/frobt.2021.768092>
- Gómez De Ágreda, Á. (2020). Ethics of autonomous weapons systems and its applicability to any AI systems. *Telecommunications Policy*, 44(6), 101953. <https://doi.org/10.1016/j.telpol.2020.101953>
- Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Proceedings of the 52nd Hawaii International Conference on System Sciences.
- Ha, Y. J. (2022). South Korean Public Value Coproduction Towards‘AI for Humanity’: A Synergy of Sociocultural Norms and Multistakeholder Deliberation in Bridging the Design and Implementation of National AI Ethics Guidelines. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 267–277. <https://doi.org/10.1145/3531146.3533091>
- Habermas, J. (1984). *The Theory of Communicative Action: Reason and the Rationalization of Society* (1st ed.). Beacon Press.

- Habermas, J. (1993). *Justification and application: Remarks on discourse ethics*. MIT Press.
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hauer, T. (2020). Machine Ethics, Allosterity and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions? *Society*, 57(4), 425–433. <https://doi.org/10.1007/s12115-020-00506-2>
- Hauer, T. (2022). Incompleteness of moral choice and evolution towards fully autonomous AI. *Humanities and Social Sciences Communications*, 9(1), 38. <https://doi.org/10.1057/s41599-022-01060-4>
- Hawkins, W., & Mittelstadt, B. (2023). The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 261–270. <https://doi.org/10.1145/3593013.3593995>
- Henderson, T. (2019). Teaching Data Ethics: We’re going to ethics the heck out of this. *Proceedings of the 3rd Conference on Computing Education Practice*, 1–4. <https://doi.org/10.1145/3294016.3294017>
- Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3), 101772. <https://doi.org/10.1016/j.jsis.2023.101772>
- Hindocha, S., & Badea, C. (2022). Moral exemplars for the virtuous machine: The clinician’s role in ethical artificial intelligence for healthcare. *AI and Ethics*, 2(1), 167–175. <https://doi.org/10.1007/s43681-021-00089-6>
- Höffe, O. (Ed.). (2013). *Einführung in die utilitaristische Ethik: Klassische und zeitgenössische Texte* (5., überarbeitete und erweiterte Auflage). A. Francke Verlag.
- Hooker, J. N., & Kim, T. W. N. (2018). Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. *Proceedings of the 2018 AAI/ACM Conference on AI, Ethics, and Society*, 130–136. <https://doi.org/10.1145/3278721.3278753>
- Horneber, D., & Laumer, S. (2023). Algorithmic Accountability. *Business & Information Systems Engineering*, 65(6), 723–730. <https://doi.org/10.1007/s12599-023-00817-8>
- Horváth, I. (2022). AI in interpreting: Ethical considerations. *Across Languages and Cultures*, 23, 1–13. <https://doi.org/10.1556/084.2022.00108>
- Houlgate, S. (2006). *The opening of Hegel’s logic: From being to infinity*. Purdue University Press.

- Islam, G., & Greenwood, M. (2021). Reconnecting to the Social in Business Ethics. *Journal of Business Ethics*, 170(1), 1–4. <https://doi.org/10.1007/s10551-021-04775-7>
- Jantunen, M., Halme, E., Vakkuri, V., Kemell, K.-K., Rebekah, R., Mikkonen, T., Nguyen Duc, A., & Abrahamsson, P. (2021). Building a Maturity Model for Developing Ethically Aligned AI Systems. *Papers of the 44th Information Systems Research Seminar in Scandinavia*. Papers of the 44th Information Systems Research Seminar in Scandinavia.
- Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics*, 65(2), 245–263. <https://doi.org/10.1007/s11192-005-0270-7>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Jones, C., Parker, M., & Bos, R. T. (2005). *Business Ethics. A Critical Approach*. Routledge Taylor & Francis Group.
- Kagan, S. (1992). The Structure of Normative Ethics. *Philosophical Perspectives*, 6, 223. <https://doi.org/10.2307/2214246>
- Kagan, S. (2018). *Normative Ethics*. Routledge. <https://doi.org/10.4324/9780429498657>
- Kago, K., & Venkataraman, P. (2023). Possibility of conjunction between altruism and egoism. *Humanities and Social Sciences Communications*, 10(1), 625. <https://doi.org/10.1057/s41599-023-02163-2>
- Kant, I. (1991). *The metaphysics of morals*. Cambridge University Press.
- Kant, I., Gregor, M. J., & Timmermann, J. (2012). *Groundwork of the metaphysics of morals* (Revised edition). Cambridge University Press.
- Katznelson, G., & Gerke, S. (2021). The need for health AI ethics in medical school education. *Advances in Health Sciences Education*, 26(4), 1447–1458. <https://doi.org/10.1007/s10459-021-10040-3>
- Kenny, L. M., Nevin, M., & Fitzpatrick, K. (2021). Ethics and standards in the use of artificial intelligence in medicine on behalf of the Royal Australian and New Zealand College of Radiologists. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 486–494. <https://doi.org/10.1111/1754-9485.13289>
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. <https://doi.org/10.1002/asi.5090140103>
- Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2022, November 15). AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers.

- Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*. Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering. <http://arxiv.org/abs/2207.01493>
- Kieslich, K., Keller, B., & Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1), 205395172210929. <https://doi.org/10.1177/20539517221092956>
- Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018). A Computational Model of Commonsense Moral Decision Making. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 197–203. <https://doi.org/10.1145/3278721.3278770>
- Knox, J. (2023). (Re)politicising data-driven education: From ethical principles to radical participation. *Learning, Media and Technology*, 48(2), 200–212. <https://doi.org/10.1080/17439884.2022.2158466>
- Koebis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6), Article 6. <https://doi.org/10.1038/s41562-021-01128-2>
- Koniakou, V. (2023). From the “rush to ethics” to the “race for governance” in Artificial Intelligence. *Information Systems Frontiers*, 25(1), 71–102. <https://doi.org/10.1007/s10796-022-10300-6>
- Kornilae, L. (2021). Kant’s doctrine of education and the problem of artificial intelligence. *Journal of Philosophy of Education*, 55(6), 1072–1080. <https://doi.org/10.1111/1467-9752.12608>
- Kowch, E. G. (2019). Ethics to Prepare Educators for Professional Service Robots in Classrooms. *2019 International Joint Conference on Information, Media and Engineering (IJCIME)*, 478–484. <https://doi.org/10.1109/IJCIME49369.2019.00102>
- Krögel, S., Ostermaier, A., & Uhl, M. (2023). *The moral authority of ChatGPT* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2301.07098>
- Kučera, D. (2019). Artificial Intelligence in the Context of Managerial Responsibility. *IMES Proceedings of the 7th International Conference: Innovation Management, Entrepreneurship and Sustainability*. International Conference: Innovation Management, Entrepreneurship and Sustainability.
- Lara, F. (2021). Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates? *Science and Engineering Ethics*, 27(4), 42. <https://doi.org/10.1007/s11948-021-00318-5>

- Leonardi. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS Quarterly*, 35(1), 147. <https://doi.org/10.2307/23043493>
- Leonardi, P. M. (2012). Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How Are They Different? Do We Need Them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and Organizing* (pp. 24–48). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199664054.003.0002>
- Li, O. (2021). Problems with “Friendly AI.” *Ethics and Information Technology*, 23(3), 543–550. <https://doi.org/10.1007/s10676-021-09595-x>
- Lim, J. H., & Kwon, H. Y. (2021). A Study on the Modeling of Major Factors for the Principles of AI Ethics. *DG.O*, 8.
- Lindkvist, L., & Llewellyn, S. (2003). Accountability, responsibility and organization. *Scandinavian Journal of Management*, 19(2), 251–273. [https://doi.org/10.1016/S0956-5221\(02\)00027-1](https://doi.org/10.1016/S0956-5221(02)00027-1)
- Lloyd, D. (1985). FRANKENSTEIN’S CHILDREN: ARTIFICIAL INTELLIGENCE AND HUMAN VALUE. *Metaphilosophy*, 16(4), 307–318.
- Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973–1986. <https://doi.org/10.1002/asi.22628>
- Mabaso, B. A. (2021). Computationally rational agents can be moral agents. *Ethics and Information Technology*, 23(2), 137–145. <https://doi.org/10.1007/s10676-020-09527-1>
- MacInnis, D. J. (2011). A Framework for Conceptual Contributions in Marketing. *Journal of Marketing*, 75(4), 136–154. <https://doi.org/10.1509/jmkg.75.4.136>
- Mackie, J. (1990). *Ethics: Inventing right and wrong*. Penguin.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Malcolm, N. (1996). *Thomas Hobbes: Leviathan*.
- Manna, R., & Nath, R. (2021). Kantian Moral Agency and the Ethics of Artificial Intelligence. *Problemos*, 100, 139–151. <https://doi.org/10.15388/Problemos.100.11>

- Mäntymäki, M., Minkkinen, M., Zimmer, M. P., & Viljanen, M. (2023). Designing an AI governance framework: From research-based premises to meta-requirements. *European Conference on Information Systems (ECIS)*.
- Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, 160(4), 835–850. <https://doi.org/10.1007/s10551-018-3921-3>
- Martinho, A., Kroesen, M., & Chorus, C. (2021). Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence. *Minds and Machines*, 31(2), 215–237. <https://doi.org/10.1007/s11023-021-09556-9>
- Mayer, A.-S., Haimerl, A., Strich, F., & Fiedler, M. (2021). How corporations encourage the implementation of AI ethics. *Twenty-Ninth European Conference on Information Systems*.
- Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433–443. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<433::AID-ASI11>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<433::AID-ASI11>3.0.CO;2-Q)
- McDonald, N., & Pan, S. (2020). Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–19. <https://doi.org/10.1145/3415218>
- McGrath, S. K., & Whitty, S. J. (2018). Accountability and responsibility defined. *International Journal of Managing Projects in Business*, 11(3), 687–707. <https://doi.org/10.1108/IJMPB-06-2017-0058>
- Meijer, A., Wiarda, M., Doorn, N., & Van De Kaa, G. (2023). Towards responsible standardisation: Investigating the importance of responsible innovation for standards development. *Technology Analysis & Strategic Management*, 1–15. <https://doi.org/10.1080/09537325.2023.2225108>
- Mill, J. S., Bentham, J., & Austin, J. (1989). *Utilitarianism* (M. Warnock, Ed.; 19. impr). Fontana Press.
- Millar, J. (2016). An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars. *Applied Artificial Intelligence*, 30(8), 787–809. <https://doi.org/10.1080/08839514.2016.1229919>
- Minkkinen, M., & Mäntymäki, M. (2023). The institutional logics underpinning organizational AI governance practices. *Scandinavian Conference on Information Systems (SCIS)*.

- Mirbabaie, M., Brendel, A. B., Faculty of Business and Economics, Technische Universität Dresden, Hofeditz, L., & Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1), 726–753. <https://doi.org/10.17705/1CAIS.05034>
- Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE INTELLIGENT SYSTEMS*.
- Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, 31(2), 239–256. <https://doi.org/10.1007/s11023-021-09563-w>
- Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. (2020). The Ethical Implications of Using Artificial Intelligence in Auditing. *Journal of Business Ethics*, 167(2), 209–234. <https://doi.org/10.1007/s10551-019-04407-1>
- Mutch, A. (2013). Sociomateriality—Taking the wrong turning? *Information and Organization*, 23(1), 28–40. <https://doi.org/10.1016/j.infoandorg.2013.02.001>
- Neubert, M. J., & Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. *Business Horizons*, 63(2), 195–204. <https://doi.org/10.1016/j.bushor.2019.11.001>
- Nevanperä, M., Rajamäki, J., & Helin, J. (2021). Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers. *Procedia Computer Science*, 192, 2330–2339. <https://doi.org/10.1016/j.procs.2021.08.223>
- Niederman, F., & Baker, E. W. (2023). Ethics and AI Issues: Old Container with New Wine? *Information Systems Frontiers*, 25(1), 9–28. <https://doi.org/10.1007/s10796-022-10305-1>
- Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-023-01635-y>
- Nowak, E. (2018). Can human and artificial agents share an autonomy, categorical imperative-based ethics and “moral” selfhood? *Filozofia Publiczna i Edukacja Demokratyczna*, 6(2), 169–208. <https://doi.org/10.14746/fped.2017.6.2.20>
- Obi, I., & Gray, C. M. (2023). Auditing Practitioner Judgment for Algorithmic Fairness Implications. *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 01–05. <https://doi.org/10.1109/ETHICS57328.2023.10154992>

- Ong, D. C. (2021). An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. <https://doi.org/10.1109/ACII52823.2021.9597441>
- Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. *Organization Studies*, 28(9), 1435–1448. <https://doi.org/10.1177/0170840607081138>
- Orlikowski, W. J. (2010). The sociomateriality of organisational life: Considering technology in management research. *Cambridge Journal of Economics*, 34(1), 125–141. <https://doi.org/10.1093/cje/bep058>
- Palmer, A., & Schwan, D. (2024). More Process, Less Principles: The Ethics of Deploying AI and Robotics in Medicine. *Cambridge Quarterly of Healthcare Ethics*, 33(1), 121–134. <https://doi.org/10.1017/S0963180123000087>
- Pansoni, S., Tiribelli, S., Paolanti, M., Di Stefano, F., Frontoni, E., Malinverni, E. S., & Giovanola, B. (2023). ARTIFICIAL INTELLIGENCE AND CULTURAL HERITAGE: DESIGN AND ASSESSMENT OF AN ETHICAL FRAMEWORK. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2–2023, 1149–1155. <https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-1149-2023>
- Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023). Responsible Digital Transformation for a Sustainable Society. *Information Systems Frontiers*, 25(3), 945–953. <https://doi.org/10.1007/s10796-023-10406-5>
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Parlangeli, O., Guidi, S., Marchigiani, E., Palmitesta, P., Andreadis, A., & Roncato, S. (2020). How guilty is a robot who kills other robots? *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–5. <https://doi.org/10.1109/IISA50023.2020.9284338>
- Passlack, N., Heyder, T., Klemm, F., & Posegga, O. (2023). How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success. *Forty-Fourth International Conference on Information Systems*. Forty-Fourth International Conference on Information Systems, Hyderabad, India.
- Polyviou, A., & Zamani, E. D. (2022). TOWARDS EUROPE’S AI STRATEGY: DESIRES & REALITIES. *Thirtieth European Conference on Information Systems*.

- Pruski, M. (2023). Ethics framework for predictive clinical AI model updating. *Ethics and Information Technology*, 25(3), 48. <https://doi.org/10.1007/s10676-023-09721-x>
- Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*, 9(1), 205395172210806. <https://doi.org/10.1177/20539517221080676>
- Rai, A., Constantinides, P., & Sarker, S. (2019). Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, 43(1), iii–ix.
- Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021). You Can’t Sit With Us: Exclusionary Pedagogy in AI Ethics Education. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 515–525. <https://doi.org/10.1145/3442188.3445914>
- Rawls, J. (1971). *A Theory of Justice*. Cambridge (Mass.).
- Rezaee, Z. (2009). *Corporate governance and ethics*. John Wiley & Sons.
- Rismani, S., & Moon, Aj. (2023). What does it mean to be a responsible AI practitioner: An ontology of roles and skills. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 584–595. <https://doi.org/10.1145/3600211.3604702>
- Roberts, H., Cows, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2023). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, 39(2), 79–97. <https://doi.org/10.1080/01972243.2022.2124565>
- Roberts, J. S., & Montoya, L. N. (2022). Contextualizing Artificially Intelligent Morality: A Meta-Ethnography of Top-Down, Bottom-Up, and Hybrid Models for Theoretical and Applied Ethics in Artificial Intelligence. *Future of Information and Communication Conference*, arXiv:2204.07612. <http://arxiv.org/abs/2204.07612>
- Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, 33(1), 100925. <https://doi.org/10.1016/j.hrmr.2022.100925>
- Rousseau, J. J. (1964). *The social contract (1762)*. Londres.
- Saariluoma, P., & Leikas, J. (2020). Designing Ethical AI in the Shadow of Hume’s Guillotine. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), *Intelligent Human Systems Integration 2020* (Vol. 1131, pp. 594–599). Springer International Publishing. https://doi.org/10.1007/978-3-030-39512-4_92

- Sætra, H. S., & Danaher, J. (2023). Resolving the battle of short- vs. Long-term AI risks. *AI and Ethics*.
<https://doi.org/10.1007/s43681-023-00336-y>
- Schlimbach, R., & Khosrawi-Rad, B. (2022). Towards Ethical Design Features for Pedagogical Conversational Agents. *Americas Conference on Information Systems (AMCIS)*.
- Schmager, S., Pappas, I., & Vassilakopoulou, P. (2023). Defining Human-Centered AI: a Comprehensive Review of HCAI Literature. *The 15th Mediterranean Conference on Information Systems (MCIS)*.
- Schneider, J., Meske, C., & Bikic, A. (2023). How individuals can shape AI through data—An AI literacy and morality perspective. *Thirty-First European Conference on Information Systems*.
- Seo, H., & Thorson, S. (2022). Computation, Rule Following, and Ethics in AIs. *Proceedings of the 55th Hawaii International Conference on System Sciences*. Proceedings of the 55th Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2022.132>
- Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021). From Ethical AI Principles to Governed AI. *Forty-Second International Conference on Information Systems*.
- Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7(1), 119.
<https://doi.org/10.1057/s41599-020-00614-8>
- Seymour, M. (2018). Artificial Intelligence Is No Match for Human Stupidity: Ethical Reflections on Avatars and Agents. *Australasian Conference on Information Systems 2018*. <https://doi.org/10.5130/acis2018.db>
- Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22(5), 648–663. <https://doi.org/10.1080/1369118X.2019.1568515>
- Shevlin, H. (2021). How Could We Know When a Robot was a Moral Patient? *Cambridge Quarterly of Healthcare Ethics*, 30(3), 459–471. <https://doi.org/10.1017/S0963180120001012>
- Shih, P.-K., Lin, C.-H., Wu, L. Y., & Yu, C.-C. (2021). Learning Ethics in AI—Teaching Non-Engineering Undergraduates through Situated Learning. *Sustainability*, 13(7), 3718.
<https://doi.org/10.3390/su13073718>
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, 109–124. <https://doi.org/10.17705/1thci.00131>

- Siapka, A. (2022). Towards a Feminist Metaethics of AI. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 665–674. <https://doi.org/10.1145/3514094.3534197>
- Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, 31(2), 74–87. <https://doi.org/10.4018/JDM.2020040105>
- Simpson, P. (1992). Contemporary Virtue Ethics and Aristotle. *The Review of Metaphysics*, 45(3), 503–524.
- Singer, P. (1968). *Applied Ethics*. Oxford University Press.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Smith, N., & Vickers, D. (2021). Statistically responsible artificial intelligences. *Ethics and Information Technology*, 23(3), 483–493. <https://doi.org/10.1007/s10676-021-09591-1>
- Søvik, A. O. (2022). How a non-conscious robot could be an agent with capacity for morally responsible behaviour. *AI and Ethics*, 2(4), 789–800. <https://doi.org/10.1007/s43681-022-00140-0>
- Stahl, B. C., Brooks, L., Hatzakis, T., Santiago, N., & Wright, D. (2023). Exploring ethics and human rights in artificial intelligence – A Delphi study. *Technological Forecasting and Social Change*, 191, 122502. <https://doi.org/10.1016/j.techfore.2023.122502>
- Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. <https://doi.org/10.1016/j.ijinfomgt.2023.102700>
- Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, 45, 105661. <https://doi.org/10.1016/j.clsr.2022.105661>
- Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral Judgments in the Age of Artificial Intelligence. *Journal of Business Ethics*, 178(4), 917–943. <https://doi.org/10.1007/s10551-022-05053-w>
- Sun, F., & Ye, R. (2023). Moral Considerations of Artificial Intelligence. *Science & Education*, 32(1), 1–17. <https://doi.org/10.1007/s11191-021-00282-3>
- Susser, D. (2019). Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 403–408. <https://doi.org/10.1145/3306618.3314286>

- Swanepoel, D. (2021). The possibility of deliberate norm-adherence in AI. *Ethics and Information Technology*, 23(2), 157–163. <https://doi.org/10.1007/s10676-020-09535-1>
- Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, 40(2), 137–157. <https://doi.org/10.1080/14494035.2021.1928377>
- Tajalli, P. (2021). AI Ethics and the Banality of Evil. *Ethics and Information Technology*, 23, 447–454. <https://doi.org/10.1007/s10676-021-09587-x>
- Tasioulas, J. (2018). First Steps Towards an Ethics of Robots and Artificial Intelligence. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3172840>
- Taylor, A. M. (1979). Dialectics of nature or dialectics of practice? *Dialectical Anthropology*, 289–308.
- Telkamp, J. B., & Anderson, M. H. (2022). The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. *Journal of Business Ethics*, 178(4), 961–976. <https://doi.org/10.1007/s10551-022-05057-6>
- Templier, M., & Paré, G. (2018). Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, 27(5), 503–550. <https://doi.org/10.1080/0960085X.2017.1398880>
- Terzis, P. (2020). Onward for the freedom of others: Marching beyond the AI ethics. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3351095.3373152>
- Tidjon, L. N., & Khomh, F. (2023). The Different Faces of AI Ethics Across the World: A Principle-to-Practice Gap Analysis. *IEEE Transactions on Artificial Intelligence*, 4(4), 820–839. <https://doi.org/10.1109/TAI.2022.3225132>
- Tóth, Z., Caruana, R., Gruber, T., & Loebbecke, C. (2022). The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability. *Journal of Business Ethics*, 178(4), 895–916. <https://doi.org/10.1007/s10551-022-05050-z>
- Townsend, B., Paterson, C., Arvind, T. T., Nemirovsky, G., Calinescu, R., Cavalcanti, A., Habli, I., & Thomas, A. (2022). From Pluralistic Normative Principles to Autonomous-Agent Rules. *Minds and Machines*, 32(4), 683–715. <https://doi.org/10.1007/s11023-022-09614-w>

- Unver, M. B. (2023). Reebuilding ‘ethics’ to Govern AI: How to re-set the boundaries for the legal sector? *International Conference on Artificial Intelligence and Law (ICAIL)*.
<https://doi.org/10.1145/3594536.3595156>
- Upton, C. L. (2009). Virtue Ethics and Moral Psychology: The Situationism Debate. *The Journal of Ethics*, 13(2–3), 103–115. <https://doi.org/10.1007/s10892-009-9054-2>
- Üsdiken, B., & Pasadeos, Y. (1995). Organizational Analysis in North America and Europe: A Comparison of Co-citation Networks. *Organization Studies*, 16(3), 503–526.
<https://doi.org/10.1177/017084069501600306>
- Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2019). Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study. In X. Franch, T. Männistö, & S. Martínez-Fernández (Eds.), *Product-Focused Software Process Improvement* (Vol. 11915, pp. 331–338). Springer International Publishing.
https://doi.org/10.1007/978-3-030-35333-9_24
- Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software*, 182, 111067.
<https://doi.org/10.1016/j.jss.2021.111067>
- Van De Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, 30(3), 385–409. <https://doi.org/10.1007/s11023-020-09537-4>
- Van De Voort, M., Pieters, W., & Consoli, L. (2015). Refining the ethics of computer-made decisions: A classification of moral mediation by ubiquitous machines. *Ethics and Information Technology*, 17(1), 41–56. <https://doi.org/10.1007/s10676-015-9360-2>
- Vanderelst, D., & Winfield, A. (2018). The Dark Side of Ethical Robots. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 317–322. <https://doi.org/10.1145/3278721.3278726>
- Vanhée, L., & Borit, M. (2022). Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence. *Journal of Artificial Intelligence Research*, 73, 619–631.
<https://doi.org/10.1613/jair.1.13135>
- Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scandinavian Journal of Management*, 34(2).
- Vică, C., Voinea, C., & Uszkai, R. (2021). The Emperor is Naked: Moral Diplomacies and the Ethics of AI. *Információs Társadalom*, XXI(2), 83–96.

- Vidu, C., Zbucea, A., Mocanu, R., & Pinzaru, F. (2020). ARTIFICIAL INTELLIGENCE AND THE ETHICAL USE OF KNOWLEDGE. *Strategica*, 773–784.
- Volkman, R., & Gabriels, K. (2023). AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement. *Science and Engineering Ethics*, 29(2), 11. <https://doi.org/10.1007/s11948-023-00428-2>
- Walke, F., Bennek, L., & Winkler, T. J. J. (2023). AI in Government: A Study on Explainability of High-Risk AI-Systems in Law Enforcement & Police Service. *Wirtschaftsinformatik (WI)*.
- Wang, Y., Xiong, M., & Olya, H. (2020). Toward an Understanding of Responsible Artificial Intelligence Practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Proceedings of the 53rd Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2020.610>
- Weaver, G. R., & Trevino, L. K. (1994). Normative And Empirical Business Ethics: Separation, Marriage Of Convenience, Or Marriage Of Necessity? *Business Ethics Quarterly*, 4(2), 129–143. <https://doi.org/10.2307/3857485>
- Weber, R. H. (2020). Socio-ethical values and legal rules on automated platforms: The quest for a symbiotic relationship. *Computer Law & Security Review*, 36, 105380. <https://doi.org/10.1016/j.clsr.2019.105380>
- Wehrens, R., Stevens, M., Kostenzer, J., Weggelaar, A. M., & De Bont, A. (2023). Ethics as Discursive Work: The Role of Ethical Framing in the Promissory Future of Data-driven Healthcare Technologies. *Science, Technology, & Human Values*, 48(3), 606–634. <https://doi.org/10.1177/01622439211053661>
- Weiss, A., Vrecar, R., Zamiechowska, J., & Purgathofer, P. (2021). Using the Design of Adversarial Chatbots as a Means to Expose Computer Science Students to the Importance of Ethics and Responsible Design of AI Technologies. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (Vol. 12934, pp. 331–339). Springer International Publishing. https://doi.org/10.1007/978-3-030-85613-7_24
- Westerstrand, S. (2023). Ethics in the intersection of AI and democracy: The AIDEM Framework. *Thirty-First European Conference on Information Systems*.
- Wight, J. B. (2014). Economics within a Pluralist Ethical Tradition. *Review of Social Economy*, 72(4), 417–435. <https://doi.org/10.1080/00346764.2014.960661>

- Willems, J., Schmidhuber, L., Vogel, D., Ebinger, F., & Vanderelst, D. (2022). Ethics of robotized public services: The role of robot design and its actions. *Government Information Quarterly*, 39(2), 101683. <https://doi.org/10.1016/j.giq.2022.101683>
- Williams, R., Kaputsos, S. P., & Breazeal, C. (2021). Teacher Perspectives on How To Train Your Robot: A Middle School AI and Ethics Curriculum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15678–15686. <https://doi.org/10.1609/aaai.v35i17.17847>
- Wong, A. (2021). Ethics and Regulation of Artificial Intelligence. In E. Mercier-Laurent, M. Ö. Kayalica, & M. L. Owoc (Eds.), *Artificial Intelligence for Knowledge Management* (Vol. 614, pp. 1–18). Springer International Publishing. https://doi.org/10.1007/978-3-030-80847-1_1
- Wright, S. A. (2020). AI in the Law: Towards Assessing Ethical Risks. *2020 IEEE International Conference on Big Data (Big Data)*, 2160–2169. <https://doi.org/10.1109/BigData50022.2020.9377950>
- Wu, W., Huang, T., & Gong, K. (2020). Ethical Principles and Governance Technology Development of AI in China. *Engineering*, 6(3), 302–309. <https://doi.org/10.1016/j.eng.2019.12.015>
- Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building Ethics into Artificial Intelligence. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 5527–5533. <https://doi.org/10.24963/ijcai.2018/779>
- Zebrowski, R. (2021). Artificial Instinct: Lem’s Robots as a Model Case for AI. *Pro-Fil*, 92. <https://doi.org/10.5817/pf21-3-2423>
- Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *Journal of Artificial Intelligence Research*, 71. <https://doi.org/10.1613/jair.1.12895>
- Zhang, J., & Zhang, Z. (2023). Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*, 23(1), 7. <https://doi.org/10.1186/s12911-023-02103-9>
- Zhang, X., Lee, S. K., Kim, W., & Hahn, S. (2023). “Sorry, it was my fault”: Repairing trust in human-robot interactions. *International Journal of Human-Computer Studies*, 175, 103031. <https://doi.org/10.1016/j.ijhcs.2023.103031>
- Zupic, I., & Čater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods*, 18(3), 429–472. <https://doi.org/10.1177/1094428114562629>

APPENDICES

Appendix I. Ethical Views

Table 5 Ethical views

Ethical Views	Definition and Emphasis	Foundational Philosophers
Normative ethics	Normative ethics aims to establish standards or norms for determining what is morally right or wrong (Brendel et al., 2021). It involves the formulation and justification of moral principles and guidelines that can be used to guide ethical decision-making. This branch of ethics seeks to provide answers to moral questions, such as “What should I do?” or “What is the right course of action?” Normative ethics is concerned with developing ethical theories, such as deontology, consequentialism, and virtue ethics, which offer frameworks for evaluating moral conduct (Brendel et al., 2021).	Kagan, S. (1992, 2018). <i>Normative ethics</i> . Routledge.
Meta-ethics	Meta-ethics explores the nature, foundation, and meaning of ethical concepts, focusing on questions about the origins and status of moral values. It addresses issues such as whether moral truths exist independently of human beliefs, how moral judgments can be justified, and whether morality is objective or subjective. This branch of ethics does not prescribe specific actions but rather analyzes the philosophical, epistemological, and ontological nature of ethics and morality (Tajalli, 2021).	Mackie, J. L. (1990). <i>Ethics: Inventing right and wrong</i> . Penguin.
Descriptive ethics	Descriptive ethics involves the empirical study of people’s moral beliefs, practices, and behaviors (Weaver & Trevino, 1994). It seeks to understand how individuals and societies perceive moral issues and how these perceptions influence their actions (Weaver & Trevino, 1994). This approach is often used in sociology, anthropology, and psychology to examine moral diversity and identify patterns in moral thinking across different cultures and communities. Descriptive ethics does not judge or evaluate moral beliefs but seeks to describe and explain them as they exist (Weaver & Trevino, 1994).	Gert, B. (2005). <i>Morality: Its nature and justification</i> . Oxford University Press.
Applied ethics	Applied ethics involves the practical application of ethical theories and principles to real-world issues and dilemmas (Brendel et al., 2021). It examines specific ethical problems in various domains, such as medicine, business, law, and environmental ethics, and seeks to provide guidance on how to address these issues. Applied ethics bridges the gap between theoretical ethical concepts and the challenges faced in everyday life, offering concrete solutions and moral reasoning for specific cases. This field is concerned with how ethical principles can be effectively implemented in diverse contexts. In the context of human-AI interaction, it encompasses AI ethics concerning the design and regulation of AI technologies based on ethical principles (ethics of AI) to enforce AI’s ethical behavior (ethical AI) (Siau & Wang, 2020), as well as business ethics as principles that guide the ethical behavior of organizational stakeholders (Floridi et al., 2018; Rezaee, 2009, p. 64). AI ethics is often related to machine ethics, focusing on moral frameworks that guide the behavior of machines themselves, especially in situations where human intervention is minimal or absent (Anderson & Anderson, 2007).	Singer, P. (Ed.). (1968). <i>Applied ethics</i> . Oxford University Press.

Appendix II. Bibliometric Word Analysis

Fig. 3 Most often connected terms (n=6,523)

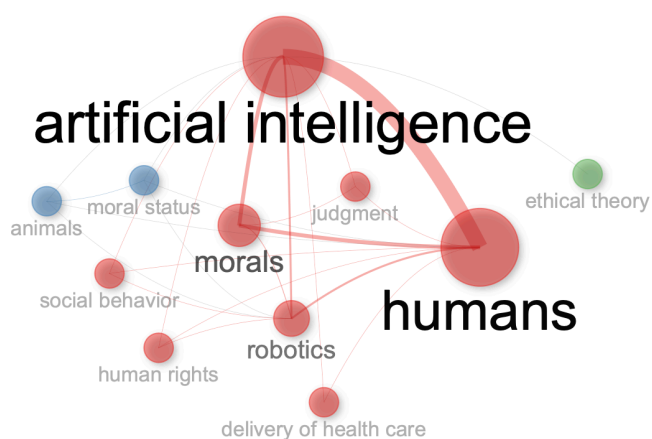


Fig. 4 Word families (n=6,523)

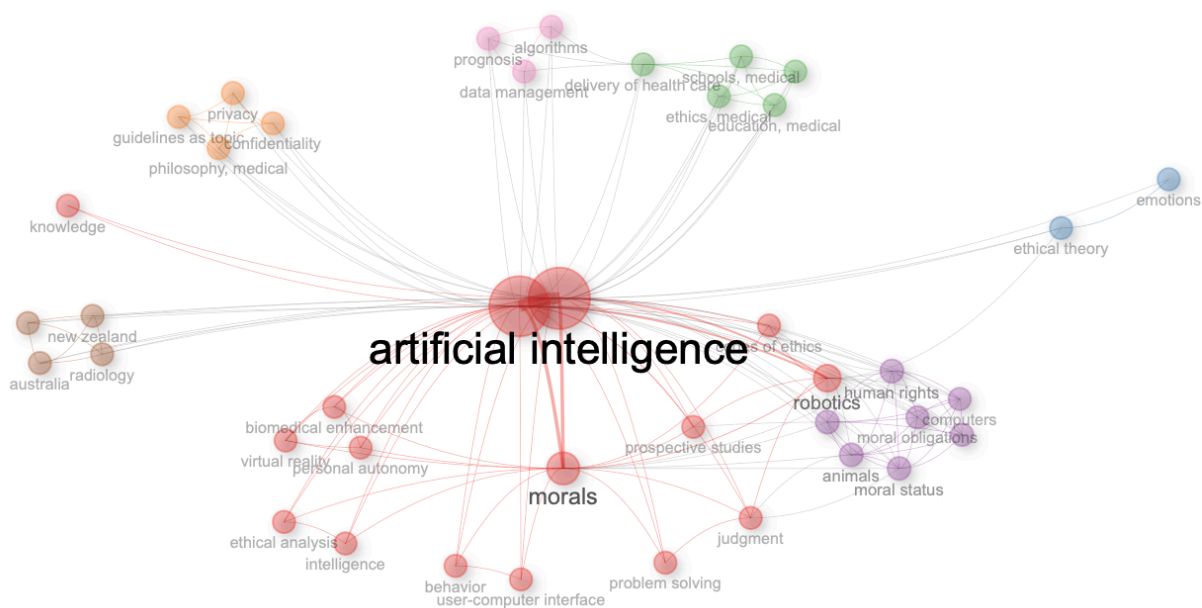
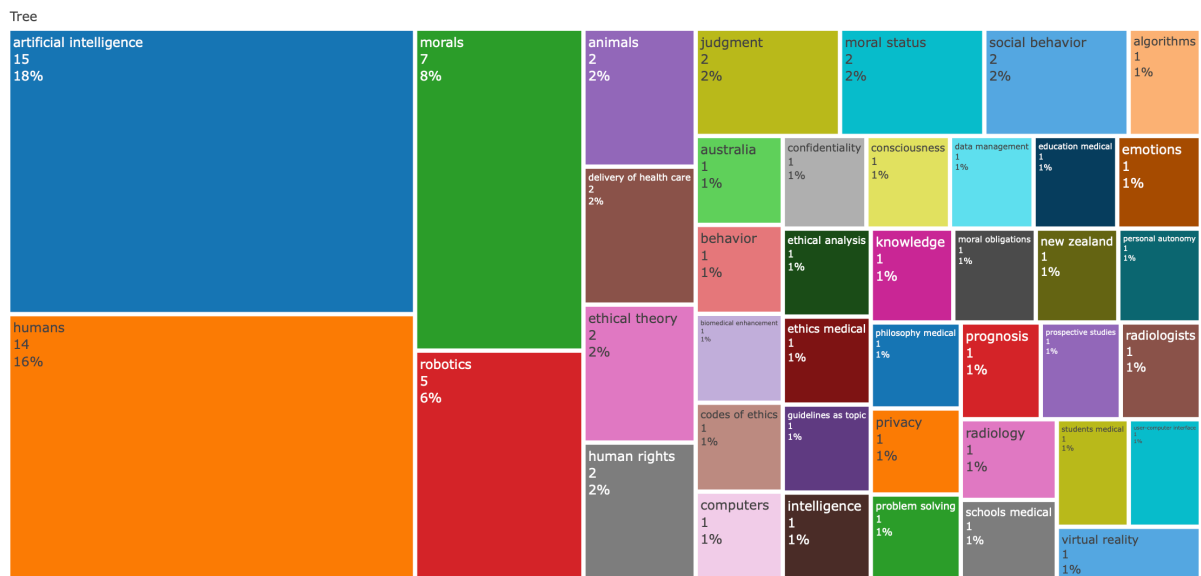


Fig. 5 Keyword appearance (n=6,523)



Appendix III. Manual Format Screening

Table 6 Format screening

Exclusion Criteria	Description	Number of Eliminated Articles
Non-scientific articles (n= 612 articles)	Gray literature, book chapters, book reviews, dissertations, tutorials, bachelor's or master's theses, workshops, posters, letter of editors, special issues, symposiums, talks, magazine, and articles of non-scientific conferences.	• ACM: 38 articles • AIS: 22 articles • BSC: 219 articles • IEEE: 85 articles • JSTOR: 3 articles • OVID: 41 articles • Science Direct: 0 articles • Web of Science: 204 articles
Incomplete articles (n= 40 articles)	Short papers or research-in-progress papers, extended abstracts.	• ACM: 9 articles • AIS: 6 articles • BSC: 1 article • IEEE: 11 articles • JSTOR: 0 articles • OVID: 1 article • Science Direct: 0 articles • Web of Science: 12 articles
Non-English articles (n= 89 articles)	Articles not completely written in English are excluded.	• ACM: 1 article • AIS: 0 articles • BSC: 19 articles • IEEE: 0 articles • JSTOR: 2 articles • OVID: 40 articles • Science Direct: 0 articles • Web of Science: 27 articles
Stand-alone review of academic literature (n= 79 articles)	Standalone literature review that do not contribute further to theory are excluded. This includes reviews that merely aggregate previous findings without offering new insights, extensions, or developments to existing conceptualizations, frameworks, or theories.	• ACM: 4 articles • AIS: 4 articles • BSC: 5 articles • IEEE: 5 articles • JSTOR: 1 article • OVID: 3 articles • Science Direct: 0 articles • Web of Science: 57 articles
Total: 820 articles		

Appendix IV. Relevance Screening

Table 7 Relevance screening

Exclusion Criteria	Description	Number of Eliminated Articles
Non-ethical: Articles that cannot be related to ethical theories or philosophical perspectives. (n=310 articles)	Articles that do not offer any ethical and philosophical view on ethics or morality (referring to meta-ethics, normative ethics, descriptive ethics, or applied ethics) are excluded.	• ACM: 10 articles • AIS: 37 articles • BSC: 5 articles • IEEE: 47 articles • JSTOR: 0 articles • OVID: 9 articles • Science Direct: 3 articles • Web of Science: 199 articles
No own perspective: Articles that do not offer an own perspective on ethics or moral agency from the researcher's or theorist's standpoint. (n=1,389 articles)	Articles must present an original viewpoint on ethics and morality in the context of human-AI interaction to provide dedicated contributions to our research subject. If ethics or moral agency is only addressed in the context of interviewee's statements without an interpretation of the authors, articles will be excluded. Additionally, ethical issues, dilemmas, or principles should concern the context of human-AI interaction.	• ACM: 46 articles • AIS: 113 articles • BSC: 29 articles • IEEE: 279 articles • JSTOR: 4 articles • OVID: 54 articles • Science Direct: 17 articles • Web of Science: 847 articles
No AI technology: Articles that do not involve technologies that can be aligned to AI demonstrating certain levels of intelligences and interactivity. (n=92 articles)	Articles must address technologies align with AI's evolving agency by carrying out tasks that typically require human cognition and demonstrate a certain level human-like behavior, semi-autonomous activity, and self-learning capabilities. Hence, we included articles that do not claim the term "AI" specifically, in case they provide ideas on independent and self-learning technologies.	• ACM: 0 articles • AIS: 0 articles • BSC: 1 article • IEEE: 13 articles • JSTOR: 2 articles • OVID: 30 articles • Science Direct: 2 articles • Web of Science: 44 articles
No human-AI interaction: Articles that do not focus on AI being used in several forms of human-AI interaction. (n=275 articles)	Articles must focus on human beings that somehow interact with AI. For instance, articles must involve a human part when designing, managing, using, deploying, teaching, researching AI technologies. Solely focusing on the functionalities of totally autonomous systems with no human involvement are excluded.	• ACM: 8 articles • AIS: 26 articles • BSC: 6 articles • IEEE: 56 articles • JSTOR: 0 articles • OVID: 0 articles • Science Direct: 1 article • Web of Science: 178 articles
Total: 2,066 articles		

Appendix V. Final Sample Overview and Description

Table 8 Final sample

Author/s (Year) “Title” Outlet*	Main Focus on Ethics in Human-AI Interaction	Type and Presence of Ethical Views	Perspective on Moral Agency	Research Description
Agbese et al. (2023) “Implementing AI Ethics: Making Sense of Ethical Requirements” EASE Proceedings	<i>AI Responsibility</i> This study deals with the topic of AI ethics and relates it to risk management. For example, they claim that AI systems with well-developed risk management within a company may interact with data from another agency (human or non-human) to create “new risk avenues” (p. 70). Their understanding of AI ethics is that it aims to ensure that people, processes, and organizations that engineer AI assume responsibility and are engineered ethically in line with moral codes and principles.	<i>Normative Ethics</i> The work underpins their ethical requirements based on prescribed ethical principles, which can be linked to deontology (implicitly addressed).	<i>HMA & AMA: hybrid moral agency</i> The authors claim that with respect to risk management, AI systems interact with data from another agency (human or non-human) to create new risk avenues.” (p. 70)	<i>Qualitative empirical work</i> The authors conducted interviews with ten Finish software engineering executives in middle and higher-level management. Their work has practical and design-oriented contributions.
Allen et al. (2000) “Prolegomena to any Future Artificial Moral Agent” Journal of Experimental & Theoretical Artificial Intelligence	<i>Autonomous robots</i> This study motivates the human-AI perspective because “robots possessing autonomous capacities to do things that are useful to humans will also have the capacity to do things that are harmful to humans and other sentient beings.” (p. 251)	<i>Normative Ethics</i> Three ethical views are pointed out in order to approach an artificial moral agency (deontology, consequentialism, and virtue ethics are explicitly addressed).	<i>HMA & AMA: hybrid moral agency</i> The authors introduce the Moral Turing Test, which bypasses disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human ‘interrogators cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent.” (p. 354)	<i>Philosophical essay</i> This essay discusses artificial moral agency using different perspectives.
Arnold & Schuetz (2018) “The Big Red Button is Too Late: An Alternative Model for Ethical Evaluation of AI Systems” Ethics Information Technology	<i>Human as a designer</i> The study argues that humans are involved in making AI “ethical” from a technical point of view. Designing ethical AI means outlining a system architecture consisting of an ethical core layer above the hardware.	<i>Applied ethics</i> The paper discusses the application of a “big red button,” meaning that human operators interrupt or divert a system while preventing the system from learning that such an intervention is a threat.	<i>HMA: shared responsibility</i> The authors discuss the idea of human beings in charge of these systems and thus, humans are in control. However, they highlight that control afterwards is too late as interventions must occur earlier when the system is learning based on human behavior.	<i>System design</i> The authors outline how to achieve ethical AI from a technical point of view (by system design).
Ashrafian (2015a) “AlonAI: A Humanitarian Law of Artificial Intelligence and Robotics” Science Engineering Ethics	<i>Robotic law for AI-AI interaction</i> The paper discusses the idea of a robotic law, which tackles the point that future AI will interact among themselves.	<i>Normative Ethics</i> The work makes use of deontological ethics and contractarianism as they further base it on ethical considerations, such as right or wrong.	<i>HMA: full ethical agents</i> The authors noted that humans have to be in charge of the moral reasoning behind the proposed law.	<i>Position paper</i> The authors elaborate on their ideas of robotic law.
Ashrafian (2015c) “Artificial Intelligence and Robot Responsibilities : Innovating Beyond Rights”	<i>Responsibility of robots</i> The paper discusses the responsibilities of robots. In the future, robot-robot-interaction will especially be the focus of research.	<i>Meta-ethics</i> The author discusses the law on ethical AI from a meta-theoretical perspective.	<i>AMA & HMA: hybrid agency</i> The author introduces the idea that a law on AI rights can help strengthen and safeguard humanity’s key ethical aspects (human moral agency). But they also talk about the concept of artificially intelligent machines as	<i>Philosophical essay</i> The work discusses ethical issues and unclear responsibilities arising with the use of AI for the military.

Science Engineering Ethics			Artificial Moral Agents (AMAs)	
Auld et al. (2022) “ <i>Governing AI through ethical standards: learning from the experiences of other private governance initiatives</i> ” Journal of European Public Policy	<i>Learning from governance initiatives</i> This study considers on how to learn from private governance initiatives to guide AI governance.	<i>Applied ethics (implicit)</i> The paper discusses AI ethics from a regulatory perspective, learning from private governance initiatives.	<i>HMA: full ethical agents and shared responsibilities</i> The study takes perspective of public-private governance and AI governance and puts responsibility to humans , such as representatives of certain states.	<i>Philosophical essay</i> The authors make suggestions on how to govern AI based on a discourse from private governance initiatives.
Badea & Artus (2022) “ <i>Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents</i> ” LNAI Proceedings	<i>AI’s intentions and morality</i> This study takes a perspective the interpretation problem on AI and on mistakes which an artificial moral agent could make and categorizes them.	<i>Meta-ethics</i> The paper uses the Wittgensteinian approach (“no rule can contain the criteria for its own application,” p. 1) to introduce how intention mistakes and instrumental mistakes can lead to interpretation problems on AI’s morality. Thereby, meta-theory is implicitly addressed.	<i>AMA: explicit moral agent</i> The authors argue that AI must include an explicit moral program distinct from its practical reasoning program; morality can be programmed in machines when they follow external values (be able to reason morally).	<i>Position paper</i> The authors demonstrate that there is an interpretation problem with AI and AI’s moral intention.
Bankins & Formosa (2023) “ <i>The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work</i> ” Journal of Business Ethics	<i>Change of the meaning of work</i> This study analyzes the role of AI in changing human work; “Overall, according to all three of our most used moral theories there are good reasons to care about meaningful work given that it respects workers’ autonomy and their ability to exercise complex skills in helping others, contributes to their wellbeing, and allows them to flourish as complex human beings.” (p. 728)	<i>Normative ethics</i> The paper integrates deontology and virtue ethics as it discusses the aspect of AI for meaningful work under these ethical views. Thereby, meaningful work refers to the perception that one’s work has worth, significance, or a higher purpose.	<i>No AMA, only HMA</i> The authors argue that there is no moral agency; human autonomy should be saved (to ensure their well-being).	<i>Philosophical essay</i> The paper discusses the role of AI at work.
Behdadi & Munthe (2020) “ <i>A Normative Approach to Artificial Moral Agency</i> ” Minds & Machines	<i>AI entities and responsibility</i> This study focuses on the issue of to what extent artificial entities should be involved in various practices that normally assume moral agency, such as the ascription of responsibility.	<i>Normative ethics</i> The paper takes a normative lens on the AMA discussion by, in particular, focusing on deontological ethics.	<i>HMA & AMA: hybrid moral agency</i> The authors argue that in any case, a hybrid moral agency is needed because either way (fully AMA or with humans), either one of these would get excluded and potentially lack moral reasoning.	<i>Philosophical essay</i> The authors illustrate a normative approach abandoning the theoretical discussions on moral agency.
Bilal et al. (2021) “ <i>Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics</i> ” ACIS Proceedings	<i>Trust in AI</i> This work studies the most important virtues to develop trust in AI (from individual perspective).	<i>Normative Ethics</i> It explicitly refers to virtue ethics theory and uses virtues as concepts.	<i>Combining HMA and AMA</i> The authors demonstrate how individuals develop moral decision skills based on their trust in how ethical the AI is (for instance, regarding data privacy), and discuss techno moral virtues by combining HMA with AMA.	<i>Mixed-method</i> This study combines quantitative and qualitative elements using Q-methodology with 39 participants.
Bleher & Braun (2023) “ <i>Reflections on Putting AI Ethics into Practice: How Three AI Ethics</i> ”	<i>Practical AI ethics</i> The paper reviews the theory-practice gap in AI ethics and produces three dimensions.	<i>Applied ethics</i> The paper integrates three reflective dimensions that can help to put AI ethics into practice.	<i>HMA: full ethical agents</i> The authors argue that as there is no AI moral agency, human moral agency is required for oversights, control, and judgment.	<i>Philosophical essay</i> This study reflects on AI governance practices.

<i>Approaches Conceptualize Theory and Practice”</i> Science and Engineering Ethics				
Bogosian (2017) “ <i>Implementation of Moral Uncertainty in Intelligent Machines</i> ” Minds & Machines	<i>Uncertain machines</i> The paper describes a framework for designing machines that deal with the sometimes competing approaches of moral philosophy, leading to uncertainties about morality in intelligent machines.	<i>Normative ethics</i> The paper explicitly considers deontological and virtue ethics.	<i>Building AMA based on HMA:</i> The author argues that AI should be designed in a way that they are uncertain about what is morally right or wrong, so in the end, the agency lies at the machine.	<i>Philosophical essay</i> The paper introduces a computational framework on moral uncertainty as there is a lack of agreement among moral philosophers on which theory should be followed.
Boyles (2018) “ <i>A Case for Machine Ethics in Modeling Human-Level Intelligent Agents</i> ” PHILPAPERS Proceedings	<i>Human-Level intelligence?</i> This study discusses the idea of technological singularity and whether greater human-level intelligence will still exist.	<i>Normative ethics</i> The paper takes a virtue ethics perspective as relating to Aristotle’s idea of the humans’ virtue forming an agent’s conceptions of what is good.	<i>AMAs as full ethical agents</i> The authors argue that AI systems are capable of moral reasoning, judgment, and decision-making.	<i>Philosophical essay</i> The work discusses that different AI researchers use different frameworks for developing AMAs.
Bryson (2018) “ <i>Patiency is not a Virtue: The Design of Intelligent Systems and Systems of Ethics</i> ” Ethics and Information Technology	<i>Societal perspectives</i> This study discusses recommendations for maintaining either human or artefact-centered societies based on how a society is forming ethics.	<i>Normative ethics</i> The paper implicitly offers a deontological perspective and addresses virtue ethics, claiming that AI ethics would be a case for normative ethics not descriptive ethics.	<i>HMA to build AMA</i> The authors argue that values generally come from humans’ evolutionary origins– humans give the values to AI.	<i>Position paper</i> The author uses a functionalist assumption, meaning that ethics is the set of behavior that maintains a society, to explain how AI systems are constructed.
Buhmann & Fieseler (2023) “ <i>Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence</i> ” Business Ethics Quarterly	<i>Governing AI</i> The study develops a framework of responsibilities for AI innovation as well as a deliberative governance approach based on democratic principles for enacting these responsibilities.	<i>Applied ethics</i> The paper takes a perspective on responsibilities while developing AI innovation but discussing the AI system itself. Thus, it can be linked to applied ethics.	<i>HMA: shared responsibility</i> Responsibilities lie by humans for oversight, control, and so on	<i>Philosophical essay</i> The authors address the challenge of balancing between AI innovation and regulation by showing how the AI industry can involve experts to make appropriate judgments of opaque AI systems.
Burema et al. (2023) “ <i>A sector-based Approach to AI Ethics: Understanding Ethical Issues of AI-related Incidents within their Sectoral Context</i> ” AIES Proceedings	<i>Stakeholder’s perspectives on AI ethics</i> The study applies a sector-specific perspective trying to understand AI ethics from the stakeholders’ view.	<i>Normative ethics</i> Their research is guided by normative principles. Thus, deontology is implicitly used.	<i>HMA: shared responsibilities</i> The authors note that there is a form of shared responsibility for different stakeholders in different industries facing similar ethical branches.	<i>Qualitative content analysis</i> The study 125 incident cases from police, education, politics, automotive, and healthcare.
Buruk et al. (2020) “ <i>A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence</i> ” Medicine Health Care and Philosophy	<i>Developing guidelines</i> This study analyzes documents with respect to developing ethical norms and guidelines for the planning, development, production, and usage of AI technology in healthcare.	<i>Normative ethics</i> The work analyzes already produced guidelines and policy documents on AI ethics (contractarianism).	<i>HMA: shared responsibility</i> The authors argue that humans need to share their responsibilities to deal with ethical issues surrounding AI and to increase AI’s trustworthiness.	<i>Qualitative content analysis</i> This work is analyzing principles in the healthcare context to determine ethical dilemmas of practical life.
Carneiro & Veloso (2022)	<i>Arising ethical problems with AI adoption</i>	<i>Normative ethics</i>	<i>HMA, no AMA</i>	<i>Syntheses</i>

“ <i>Ethics, Transparency, Fairness and the Responsibility of Artificial Intelligence</i> ” DITTE Proceedings	The study presents ethical problems arising with the widespread adoption of AI, such as “ensuring transparency, fairness, equality, explainability and avoiding bias or discrimination” as a human-centered approach (p. 1).	The work discusses AI ethics taking a regulatory perspective by implementing AI contracts and policies, which can be linked to contractarianism.	The authors are trying to solve ethical problems from a technical standpoint (referring to robotic laws).	The paper presents the regulation, methods, and techniques that can help to address ethical problems raising up with the use of AI.
Champagne & Tonkens (2023) “ <i>A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot Harm</i> ” Science and Engineering Ethics	<i>Moral answerability</i> This study produces a framework for assigning moral answerability for robotic actions to humans and robots into categories. They extend the blank check solution.	<i>Meta-ethics</i> The paper takes different meta-theoretical perspectives on ethics and discusses the blank-check solutions from these (writing about meta-ethical fictionalism or instrumentalism). <i>Normative ethics</i> The paper addresses principles to ensure moral answerability, thereby implicitly taking a deontological perspective on it.	<i>HMA: humans remain responsible</i> The authors argue that a blank check solution is required to ensure the moral answerability of autonomous robots, meaning that the person activating a robot is willing to be answerable for whatever events ensue, even if those events stem from the robot’s autonomous decision(s). This blank check needs to be done voluntarily and before bad outcomes are known.	<i>Conceptualization</i> The authors use a taxonomy of five stances to differentiate—and defend—what is known as the “blank check” proposal on how to assign moral responsibility to AI (that was originally proposed in the context of automated warfare).
Chaput et al. (2021) “ <i>A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior</i> ” AIES Proceedings	<i>Ethical behavior</i> The study refers to ethics by design pertaining to the production of ethical behavior by the agent itself through the integration of specific abilities, e.g., reasoning or learning moral rules.	<i>Normative ethics</i> This work uses a deontological perspective by claiming that AI agents can learn ethical behaviors from experience while being guided by an existing ethical framework to force constraints on it and prevent it from diverging (e.g., Kantian Categorical Imperative).	<i>Hybrid moral agency</i> The authors propose a hybrid approach, where AI is guided by some moral frameworks and by adhering to Kant’s Categorical Imperative.	<i>System design</i> This study develops a formal model and experiments on a hypothetical Smart Grid use case in electricity production.
Coeckelbergh (2020) “ <i>Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability</i> ” Science and Engineering Ethics	<i>Responsibility attribution</i> This study focuses on responsibility attribution in contexts where humans interact with AI agents.	<i>Normative ethics</i> The authors argue through virtue ethics as they use two Aristotelian conditions for responsibility.	<i>Limited HMA and humans as moral patients</i> The author discusses that decision-making is connected to responsibility, but humans’ moral agency is limited (e.g., ignorance of what AI is exactly doing). The author further argues that humans who are affected by the action of an AI require moral answerability.	<i>Philosophical essay</i> The work discusses responsibility questions of AI actions in the context of moral agency and moral patience.
Colmenarejo et al. (2022) “ <i>Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation</i> ” AIES Proceedings	<i>AI governance through different perspectives</i> This study investigates AI regulation from an interdisciplinary perspective by considering how law and industries achieve agreement with European values.	<i>Normative ethics</i> The study analyzes AI regulation from different perspectives that can be linked to various ethical views, such as deontological ethics and consequentialism.	<i>AMA should be regulated and based on HMA</i> The authors argue that European values can guide AI design and extend morality within AI systems.	<i>Position paper</i> The paper reflects on fairness and discrimination issues of AI and how these issues are addressed within AI regulations, such as the European AI Act.
Conitzer et al. (2017) “ <i>Moral Decision Making Frameworks for Artificial Intelligence</i> ” AAAI Proceedings	<i>Automating moral judgment</i> This study introduces the idea of a machine-learning approach to automating moral judgement “...would enable AI to make moral decisions that are safer, more robust, more beneficial, and acceptable to a wider range of people” (p. 4381).	<i>Normative ethics</i> The authors argue that moral decision-making starts with humans labeling what is right based on principles, which can be linked to deontology, and involving a moral component not only on rules, which can be linked to virtue ethics.	<i>HMA: shared responsibility</i> The authors noted that joint involvement of humans in decision frameworks for AI technologies means that humans need to label first what’s right and wrong morally so that ethics is given by humans.	<i>Philosophical essay</i> The authors discuss good policies for AI applications. The authors present various ways to build AI systems that have a moral component not solely based on rules.

Constantinescu et al. (2021) “ <i>Understanding Responsibility in Responsible AI. Diagnostic Virtues and the Hard Problem of Context</i> ” Ethics and Information Technology	<i>Developing moral AI</i> This study presents a framework for humans developing moral AI and how their particular understanding of moral responsibility impacts “ethics by design, ethics in design, and ethics for designers” p.1)	<i>Normative ethics</i> The study’s framework is based on Aristotelian virtue, as it plays a big role in terms of AI responsibility and human responsibility.	<i>HMA: full ethical agents</i> The authors noted that responsibility is on the human side and can be fostered bottom-up through one’s moral values (rather than virtue ethics) or top-down through principles (rather than deontology). <i>AMA: implicit ethical agents</i> The authors argue that AI technologies can demonstrate some forms of diagnostic virtue following a rule-based form of morality given through humans.	<i>Conceptualization</i> The authors highlight difficulties in assigning responsibility to the AI, the user, or the developer by building a framework of bottom-up and top-down responsibility attribution.
Cook (2023) “ <i>Robust Artificial Moral Agents and Metanormativity</i> ” AIES Proceedings	<i>Metanormativity is required for the design of ethical AI</i> This study claims that humans cannot predict in advance which meta-normative conclusions would be sufficient to achieve intelligent, ethical AI. Designing ethical AI might still not behave ethically in all situations.	<i>Meta-ethics</i> The paper discusses the possibilities of AI ethics from a meta perspective, writing about moral error theory and meta-normativity (e.g., what if AI ceases to behave ethically).	<i>AMAs: full ethical agents</i> The authors argue that AI is capable of ethical reasoning.	<i>Philosophical essay</i> The paper explores the relationship between our ignorance concerning certain meta-normative topics and the design of ethical artificial intelligence (AI).
Cooper et al. (2022) “ <i>Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning</i> ” FACCT Proceedings	<i>Responsibility attribution</i> The article discusses responsibility and accountability questions following the idea of Nissenbaum (1995) who described barriers in moral philosophy in discussing accountability in a society using computerized systems.	<i>Meta-ethics</i> The work grounds on moral philosophy by using Nissenbaum’s barriers in moral philosophy.	<i>HMA: full ethical agents</i> The authors argue that humans are accountable for actions (e.g., for bugs) and, thus, also for the ethics put into the machines.	<i>Philosophical essay</i> The paper discusses Nissenbaum’s (1995) original work by providing a grounded discussion on the barriers in moral philosophy, bringing discourse together in a relational accountability framework in the context of data-driven algorithmic systems.
Corvite et al. (2023) “ <i>Data Subjects’ Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens</i> ” ACM HCI Proceedings	<i>Emotional AI</i> The study poses that AI could benefit from being more ethical: “It may be better in getting past the common problems of discrimination in giving better readings.” These conceptions of EAI (emotional AI) show the allure towards EAI’s promises of objective and de-biased inferences to benefit data subjects by replacing human subjectivity and bias in the workplace.” (p. 124:17, statement by participant)	<i>Normative ethics</i> The paper uses a relational ethical lens (meaning that the needs of the most impacted by a technology are at the center to critically examine the implication for them; consequentialism and deontology are thereby explicitly addressed).	<i>Limited HMA</i> The authors argue that emotional AI can enhance well-being in the working environment. However, potential risks are “overshadowed” by the perceptions of the benefits (p.1). EAI harms exist, even though there are concerns about EAI’s accuracy. Hence, positivist perceptions limited humans’ moral agency.	<i>Quantitative online survey</i> The study examines the perceived benefits and risks of 395 US adults associated with emotional AI use at the workplace.
Córdova & Vicari (2022) “ <i>Practical Ethical Issues for Artificial Intelligence in Education</i> ” TECH-EDU Proceedings	<i>Designing AI curricula</i> This study proposes a solution for ethics by designing teaching and learning processes to achieve a safer and more reliable AI for education following a top-down approach using the classic beliefs, desires and intentions (BDI) model.	<i>Normative ethics</i> The work takes on a utilitarian and a deontological perspective: the deontological dimension of the agent will guide its behavior by means of ethical principles, and its utilitarian dimension will help the AMA to solve ethical dilemmas.	<i>AMA can be used to strengthen HMA</i> The authors argue that a system architecture proposing AMA following a top-down approach can help to solve ethical dilemmas and foster collaborative learning (thus contributing to HMA).	<i>Conceptualization</i> The work tries to provide a practical proposal for integrating ethics by design in AI education using a top-down approach for AMAs following the idea of value alignment.
Dodig Crnković & Cürüklü (2012)	<i>Creating artificial morality for ethical behavior</i>	<i>Normative ethics</i> The paper investigates how ethical robots	<i>HMA: full ethical agents</i> The authors argue that humans are in charge of	<i>Philosophical essay</i> This work discusses guidelines for

“ <i>Robots: Ethical by Design</i> ” Ethics Information Technology	This study claims that artificial morality should be built into future robots/softbots, with the aim of ensuring their ethically adequate behavior.	should be designed by principles and policies (deontology and contractarianism).	responsibility by noting, “Robots are and will remain in the foreseeable future dependent on human ethical scrutiny as well as on the moral and legal responsibility of humans” (p. 63).	robots’ ethics by design.
Dyrkolbotn et al. (2018) “ <i>On the Distinction between Implicit and Explicit Ethical Agency</i> ” AIES Proceedings	<i>Implicit and explicit agents</i> The study distinguishes between implicit and explicit ethical agency by stating, “...a machine can be implicitly ethical due to a lack of morally salient options: The machine cannot choose to do something unethical if it cannot choose at all. Finally, an agent can be implicitly ethical by default; it cannot to do something unethical (in the sense of violating our expectations) if its actions have no ethical impact at all” (p. 76).	<i>Normative ethics</i> The work claims that moral values are explicitly given through existing legal categories that provide boundaries by law, thus implicitly addressing contractarianism.	<i>HMA: shared responsibility</i> The authors claim human responsibility by stating, “The agent can still be highly autonomous and highly ethical, but the underlying causes of its ethical decisions must be traced back to human agency” (p. 79).	<i>Philosophical essay</i> This work defines ethical agency in machines, distinguishing between implicit and explicit ethical agency to develop moral machines based on the boundaries of existing legal categories.
Eicher et al. (2018) “ <i>Jill Watson Doesn’t Care if You’re Pregnant: Grounding AI Ethics in Empirical Studies</i> ” AIES Proceedings	<i>Ethical issues of virtual assistants</i> This study outlines ethical issues that are important in the development and deployment of virtual teaching assistants as they might not solely used for teaching scenarios by students but also for private matters. For instance, in the response of a student stating that he will soon become a father, the virtual teaching assistant Jill Watson just provides congratulations.	<i>Descriptive ethics</i> The paper implicitly takes on a descriptive perspective by showing the role of ethics when it comes to actual development in a classroom with empirical insights.	<i>HMA as ethical baseline</i> The authors highlight that since AI demonstrates some forms of ethics that might not always fit to humans’ own ethical requirements, finding a balance is important. The pregnancy context shows that it is hard to develop AI agents that demonstrate ethical responsibilities.	<i>Case study</i> The paper showcases a prototypical development of a virtual teaching assistant (called Jill Watson). Thereby, ethical issues arise with the frequently asked questions that arise in a classroom when Jill Watson is deployed for teaching purposes but used for private matters as well.
Emdad et al. (2023) “ <i>Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence</i> ” AMCIS Proceedings	<i>AI guidelines for ethical AI</i> The study concerns ethics and morality by design: They claim that ethical AI can be designed by following a well-established ethics framework.” (p. 2)	<i>Normative Ethics</i> They adopt a utilitarian framework in designing ethical AI in healthcare (consequentialism).	<i>Build AMA based on HMA</i> The authors argue that ethical AI can be designed based on ethics frameworks made, discussed, and evaluated by humans.	<i>Interviews</i> The study conducts a secondary data analysis based on interviews and survey data with AI experts.
Erman & Furendal (2022) “ <i>The Global Governance of Artificial Intelligence: Some Normative Concerns</i> ” Moral Philosophy and Politics	<i>Societal impact of AI striving for global AI governance</i> This study discusses topics around the ethical and social impact of AI on society and claims that decisions must be made in a holistic way by a global governance of AI; benefits and burdens of AI must be discussed, and to achieve justice, it requires basic institutional governance mechanisms.	<i>Normative Ethics</i> The authors propose a theoretical framework that sets up certain normative boundary conditions through principles and the dependency of political decisions, which can be linked to deontology.	<i>HMA is broader than AMA</i> The authors argue that democratic values, political entities, and decision-making in the context of AI governance should be done holistically, which requires governance mechanisms that are not limited to AI.	<i>Position paper</i> The study discusses global AI governance norms and arising normative concerns.
Estrada (2018) “ <i>Value Alignment, Fair Play, and the Rights of Service Robots</i> ” AIS Proceedings	<i>Moral behavior</i> This study addresses primary concerns of the value alignment literature and the use of the Moral Turing Test that depends on imitations of human moral behavior through AI as a “source of inspiration in developing machines that are value-aligned” (p. 5).	<i>Meta-ethics</i> The paper takes a metatheoretical perspective on the concept of the Moral Turing Test.	<i>Influence of AMA on HMA</i> The authors argue that AI’s moral behavior can impact humans because Moral Turing Tests imitate human moral behavior.	<i>Conceptualization</i> The work discusses the Moral Turing test, which is based on the imitation of human moral behavior and provides a framework of fair play that can be used to situate robot

				rights in the context of value alignment.
Etzioni & Etzioni (2017) “Incorporating Ethics into Artificial Intelligence” Journal of Business Ethics	<i>Ethical choices in AI design</i> The study claims that the challenges posed by AI-equipped machines can be addressed by the kind of ethical choices made by human beings for millennia because they make the ethical choices in the end.	<i>Meta-ethics</i> The paper implicitly takes a meta-perspective on AI ethics by considering collective sense-making of ethics and deliberation of ethics through systems feedback and learning mechanisms based on data input of users.	<i>Integrating HMA in AI is better than developing AMA</i> The authors claim that actually, humans make decisions about ethics in the end as a collective (by data input and feedback), which might be more objective than integrating single principles of developers in the AI’s design	<i>Philosophical essay</i> The paper discusses reasons scholars hold about autonomous machines needing to make ethical decisions.
Fabiano (2019) “Robotics, Intelligent Systems, Ethics and Data Protection” APPIS Proceedings	<i>Data protection</i> The work describes the three ethical approaches, including embedded ethics, a value-sensitive design as well as an ethically aligned approach. The protection of personal data as “the right to the protection of personal data shows how much it is essential the value of any personal information. Protecting personal data means to preserve a value that is the person” is in the focus of the work (p. 1).	<i>Normative ethics</i> The work refers to deontological ethics based on rules. However, the study argues that legal frameworks are not insufficient, since “laws are necessary to regulate the life of people, but the operators have to apply them considering essential principles often not written like ethics” (p. 5). <i>Applied ethics</i> The work deals with the transition from written regulations such as laws to practice.	<i>HMA: shared responsibility of humans</i> The authors argue that since moral decisions are taken by humans only, humans remain responsible. However, there is a shift to automation and replacement through AI, which somehow requires shared forms of responsibility between human users and developers of autonomous AI technologies.	<i>Philosophical essay</i> This work makes propositions for future work and deals with the relation between AI, data protection, ethics, intelligent systems, privacy, and robotics.
Farisco et al. (2020) “Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence” Science and Engineering Ethics	<i>Ethical analysis requires theoretical and practical considerations</i> The study states that ethical analysis requires both normative and practical ethical issues and theoretical ethical issues, as “the potential replacement of humans by AI in the labor market, hype concerning AI abilities, humans lose some skills due to overreliance on AI for a growing number of activities, increasing changes in human interactions, and AI’s potentially addictive design, among others.” (p. 2419). Thereby, theoretical ethical issues deal with whether core ethical concepts can be applied to biological entities and artificial entities (e.g., AI), while normative ethical issues are tied to biological entities.	<i>Meta-ethics</i> This study proposes a methodological model to reach a comprehensive ethical analysis of AI. Thereby, the definition and clarification of crucial terms, such as intelligence, is in focus and emphasized.	<i>HMA: full ethical agent</i> The authors argue that moral agency can be described as an autonomous choice and behavior of a moral agent. Human intelligence cannot be replaced by AI.	<i>Philosophical essay</i> This work deals with a methodological model to comprehensively analyze the usage of AI.
Figueras et al. (2022) “Exploring Tensions in Responsible AI in Practice. An interview Study on AI Practices in and for Swedish Public Organizations” Scandinavian Journal of Information Systems	<i>Issues in achieving responsible AI practices</i> This study refers to responsible AI, which is based on guidelines and frameworks to design responsible and ethical AI systems linked to the principles of Jobin et al. (2019).	<i>Applied ethics</i> The study argues that ethical frameworks need to be transferred into an organizational culture to be effective. <i>Normative ethics</i> The study evaluates and analyses several ethical principles including implicitly addressing consequentialism and deontology.	<i>Aligning AMA with HMA</i> The authors argue that to overcome responsibility issues in practice, AI must be designed to be trustworthy and aligned with key ethical guidelines.	<i>Interview</i> The study conducts 13 semi-structured interviews of public organizations to investigate the awareness and perception of AI practitioners on ethical issues in the AI context. Questions address current and past ethical issues of the AI system design and the everyday handling of ethical issues.

Forsyth et al. (2021) “ <i>Imagine a More Ethical AI: Using Stories to Develop Teens’ Awareness and Understanding of Artificial Intelligence and its Societal Impacts</i> ” IEEE Proceedings	<i>Awareness of AI biases</i> The work proposes an interdisciplinary AI program to increase awareness and a better understanding of AI and its societal impacts by discussing ethical dilemmas and simulations.	<i>Descriptive ethics</i> The study carries out an experiment with students to determine their perceptions and understanding of (ethical) AI.	<i>Strengthening HMA</i> The authors argue that AI outputs can be biased, wherefore teens should be better trained to be more aware of such biases when interacting with AI.	<i>Experiment</i> This work performs a classroom experiment in the context of informal and formal learning settings.
Fukuda-Parr & Gibbson (2021) “ <i>Emerging Consensus on ‘Ethical AI’: Human Rights Critique of Stakeholder Guidelines</i> ” Global Policy	<i>Social good vs. transparency: weighting ethical principles</i> This study discusses that most ethical frameworks focus on transparency instead of principles to ensure the social good, such as discrimination, inequalities, and participation.	<i>Normative ethics</i> The study discusses various ethical principles that emerged due to raising concerns about ethical issues in the context of AI. Thus, deontology is implicitly addressed.	<i>HMA: shared responsibility</i> The authors argue that since governments and civil society are responsible for developing AI standards, ethical principles are rooted in international human rights.	<i>Syntheses</i> The work analyzes 15 guidelines on ethical practices, whereby the definition of inequality and ethics should be widened.
Gahnberg (2021) “ <i>What Rules? Framing the Governance of Artificial Agency</i> ” Policy and Society	<i>AI governance framing</i> The work focuses on the environment including programming and hardware, which mainly determines the agency of AI. Thus, governance should be aware of this environment as expectations on AI agents also require specific regulation.	<i>Normative ethics: AI governance</i> AI governance can be understood as an intersubjectively recognized rule that not only defines the expectations towards an artificial agent and its characteristics but also restricts and shapes these (deontology; implicitly addressed).	<i>AMA: implicit ethical agents</i> The author argues that AMA would involve the ability to make decisions and exhibit behavior that are opaque and unpredictable, which requires specific regulation.	<i>Philosophical essay</i> This work is an academic essay looking at how an analysis of AI governance is best grounded in an understanding of AI.
Génova et al. (2023) “ <i>Machine Ethics: Do Androids Dream of Being Good People?</i> ” Science and Engineering Ethics	<i>Machine ethics</i> This work sees machine ethics as part of moral philosophy, which deals with ensuring the ethical behavior of machines with AI.	<i>Normative ethics</i> This work implicitly presents ideas that can be linked to virtue, deontological, and utilitarian ethics.	<i>AMA: implicit ethical agents</i> The authors argue that robots can be taught based on doing the morally right thing. However, ethics and ethical behavior are more than behaving according to a specific program or code of conduct.	<i>Position paper</i> This work discusses ethics, which should be integrated into machines. The focus does not lie on the technical perspective since ethics is more than simply following a code of conduct.
Gerdes (2022) “ <i>A Participatory Data-Centric Approach to AI Ethics by Design</i> ” Applied Artificial Intelligence	<i>Focus on data when designing ethical AI</i> This study suggests a data-centered, participatory approach to AI ethics by design to draw attention to ethical challenges in the development of AI systems.	<i>Normative ethics:</i> The work argues that it is not sufficient that AI systems simply prevent harm and morally wrong actions but rather to ensure that AI systems behave beneficially and contribute to societal good in a positive way, which can be linked to consequentialism.	<i>HMA: shared responsibility</i> The authors argue that humans remain responsible since developers and experts are at the center of AI’s ethical activities. The authors encourage that ethical experts, users, and developers should work closer together.	<i>Conceptualization</i> This work addresses ethical challenges when designing and developing systems with a focus on developers and experts working together.
Ghatar et al. (2023) “ <i>Practices for Responsible AI: Findings from Interviews with Experts</i> ” AMCIS Proceedings	<i>Responsible AI practices</i> This work identifies practices for explainability, fairness, and inclusiveness of AI systems.	<i>Applied ethics</i> The work investigates the extent to which the principles for responsible development and deployment of AI have been operationalized.	<i>HMA building/training AMA:</i> The authors argue that humans are in charge of integrating responsible AI practices to ensure, for example, fairness and morality.	<i>Interviews</i> This study contains empirical qualitative research, interviewing six experts working for European organizations and six working for organizations in North America to provide practical and design-oriented contributions.
Giroux et al. (2022) “ <i>Artificial</i>	<i>Machine ethics and morality</i> This work describes machine morality as dealing with the	<i>Descriptive ethics</i> This study conducts a meta-analysis of three	<i>HMA-AMA: distributed moral responsibility</i>	<i>Experiment</i> This study contains a meta-analysis of

<i>Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI</i> Journal of Business Ethics	possession and implementation of morality and, especially, the moral capacity of autonomous machines.	previous experiments to investigate the interaction with and moral concerns of humans towards technologies vs. humans.	The authors argue that agency includes the ability to plan and execute. The biggest difference between humans and AI is the ability of humans to experience emotions. Thus, responsibilities differ between humans being accountable for designing AI technologies and users utilizing those technologies.	three previous experiments with users working in the legal sector to test whether they behave more morally with other humans than when engaging with an AI technology.
Greene et al. (2019) “Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning” HICSS Proceedings	<i>Social justice with ethical AI</i> The work focuses on ethical AI and machine learning that strives for social justice since AI technologies are part of humans’ daily lives. Thereby social justice is related to seven themes for ethical AI, consisting of 1) better building, 2) design as the locus of ethical scrutiny, 3) expert oversight, 4) machine translation, 5) objectively measured, 6) stakeholder-driven legitimacy, 7) universal concerns, and 8) values-driven determinism.	<i>Applied ethics</i> The work refers to several public partnerships, ethical boards, and declarations for responsible AI.	<i>HMA: shared responsibility of specific groups of people</i> The work highlights the crucial relation between theory and practice since “poor ethics leads to bad designs” (p. 2129). This also emphasizes the high responsibility of humans compared to the algorithm behind an AI system. Thus, human agency is an important part of ethical AI design. Nevertheless, it is usually limited to a few experts who are responsible for designing, implementing, and sometimes monitoring AI. However, for each decision and information made, a human is ultimately responsible for, never the algorithm alone.	<i>Conceptualization</i> This study assesses eight ethical considerations (explicitly related to increasing social justice) of creating ethical AI. The study focuses on AI technologies created and used for governmental/public institutions.
Gómez de Ágreda (2020) “Ethics of Autonomous Weapons Systems and its Applicability to any AI Systems” Telecommunications Policy	<i>Control over AI</i> This study in the context of autonomous weapons shows that most ethical principles apply to human use of artificial intelligence systems as long as two characteristics are met: that the way algorithms work is understood and that humans retain enough control.	<i>Normative ethics</i> The paper discusses ethical guidelines from a legal perspective in the military context, focusing on the ethics of LAWS (lethal autonomous weapons systems), which can be linked to contractarianism.	<i>HMA: full ethical agents</i> The authors argue that human agency is fully retained as HMA is required to design and regulate AI by raising the question, “Should we consider it less of a breach of human dignity to allow algorithms to decide whom to kill than to have them filter the data we access and thence distort our perception of reality and our freedom to choose? If freedom and power are the basis of human dignity, then it is conceivable that the ethics on LAWS should extend also to non-lethal activities if they affect our dignity” (p. 9).	<i>Syntheses</i> This work provides a legal perspective on the use of autonomous weapons in a military context.
Ha (2022) “South Korean Public Value Coproduction Towards ‘AI for Humanity’: A Synergy of Sociocultural Norms and Multistakeholder Deliberation in Bridging the Design and Implementation of National AI Ethics Guidelines” FACCT Proceedings	<i>Legal implementation of ethical codes</i> This work deals with the question of how governmental institutions and different interest groups can contribute to the technical and legal implementation and synthesis of ethical codes of conduct, guidelines, and principles in practice.	<i>Descriptive ethics</i> The work uses interviews to identify public values and sociocultural norms.	<i>HMA: full ethical agents</i> The author highlights that human morals and norms serve as a base for ethical guidelines.	<i>Case study</i> The paper investigates the Korean “National Guidelines for Artificial Intelligence Ethics” to better understand the transition from the design of AI ethics to the implementation based on in-depth interviews with members of government, industry, academia, and citizen groups, the four key groups holding stakes in AI

				ethics development in Korea.
Hagendorff (2020) “ <i>The Ethics of AI Ethics: An Evaluation of Guidelines</i> ” Minds & Machines	<i>AI ethics</i> This study argues that AI ethics is often understood as an add-on to technical aspects. Deontological and virtue approaches should be aligned so that ethics is perceived as a way of promoting personality, changing attitudes, encouraging responsibility, and acquiring the courage to refrain from unethical behavior.	<i>Normative ethics</i> The study analyses and compares several guidelines that are linked to deontology and how this can be aligned with individuals’ specific codes of conduct linked to virtue ethics.	<i>HMA: full ethical agents</i> The author highlights that every individual should generate motivation and adopt habits to use AI in a responsible and positive manner.	<i>Syntheses</i> This work compares 22 guidelines of AI ethics. The results show that there is a lack of implementation, as deviations usually have no consequences.
Hauer (2020) “ <i>Machine Ethics, Allosterity and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions?</i> ” Society	<i>Pluralism of AI ethics</i> The study describes the challenges that AI developers face when aiming to build better moral machines. These challenges are a lack of ethical knowledge, the plurality of ethical methods, ethical dilemmas, and machine distortion. The author criticized the reliance on traditional philosophical concepts and thought experiments in AI ethics, arguing that they are often disconnected from practical and empirical considerations, leading to unproductive debates.	<i>Normative ethics</i> The author divides AI ethics research into two primary areas. The first deals with creating and applying ethical rules and standards to ensure AI systems respect fundamental rights and are technically robust, which can be linked to deontology. The second area focuses on whether and how AI systems can behave autonomously in an ethical manner, which can be linked to virtue ethics. <i>Applied ethics</i> The work challenges traditional philosophical approaches to AI ethics and proposes a more practical, tool-based perspective for understanding and developing ethically autonomous AI systems.	<i>AMA: implicit ethical agents that depend on HMA</i> The author introduces the concept of “polymorphic recursive allosterity” in AI, which involves recursive self-improvement in AI systems. He suggests that this process could lead to ethically autonomous AI, although current systems still rely heavily on human intervention for significant improvements as AI is not free, following most ethicists and philosophers.	<i>Position paper</i> The work dismisses the philosophy of mind and consciousness as “fantastic literature,” arguing that these fields do not contribute meaningfully to the discussion of AI ethics. Instead, he advocates for a practical approach that views all human theories as tools rather than definitive truths.
Hauer (2022) “ <i>Incompleteness of moral choice and evolution towards fully autonomous AI</i> ” Humanities & Social Sciences Communications	<i>Ethics of fully autonomous AI</i> The paper outlines that the development of ethics can be described as analogous to the development of language. Overall, AI ethics covers numerous sub-questions. A better and more complete version of AI ethics is therefore needed.	<i>Normative ethics</i> The study highlights that moral issues should be prevented through deontological principles. Utilitarianism can help to find an agreement “where our too general or too specific justifications cease to work” (p. 5), and virtue ethics is mentioned as a third category of ethical choices for autonomous AI.	<i>Human moral agency</i> The author argues that AI ethics covers several sub-questions, including the establishment of institutions that implement moral issues in research.	<i>Philosophical essay</i> The authors discuss AI’s full automation of tasks, including the transfer of moral competence.
Hawkins & Mittelstadt (2023) “ <i>The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices</i> ” FACCT Proceedings	<i>The gap between norms and practices</i> The study highlights that, in recent years, more and more ethical guidelines and norms have been established in the field of AI research. However, those norms often present a gap between current AI practices.	<i>Normative ethics</i> The paper presents research ethics policies that can be linked to contractarianism. <i>Applied ethics</i> The study investigates how data enrichment can close gaps in the context of crowdsourcing.	<i>HMA: shared responsibility</i> The study highlights gaps between ethics norms and practices, encouraging shared responsibilities between users and regulators.	<i>Quantitative text analysis</i> The work deals with a gap regarding ethical disclosures in research and ethical issues while working with crowdsourced workers.
Henderson (2019) “ <i>Teaching Data Ethics</i> ” CEP Proceedings	<i>AI curricula</i> This study aims to make students aware of 1) algorithmic accountability, 2) bias in AI, 3) data protection and filter bubbles, and 4)	<i>Applied ethics</i> The work presents a new module on Data Ethics and Privacy (AI curricula).	<i>HMA: training human moral agency</i> The authors argue that humans should be trained to be aware of ethical issues in the AI context.	<i>Case study</i> The study presents a new “Data Ethics & Privacy module” for students of computer

	practical mechanisms for dealing with this topic.			science in 2018 in the UK.
Hooker & Kim (2018) “ <i>Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic</i> ” AIES Proceedings	<i>Logic vs. intuition of AI ethics</i> This work aims to achieve breakthroughs in issues that cannot be resolved by human ethicists.	<i>Normative ethics</i> The study derives ethical principles to give machines a base to explain specific actions and contribute to a transparent AI, whereby deontology is explicitly addressed.	<i>HMA-AMA: distributed moral responsibility</i> The authors note that ethical principles are still the responsibility of humans (especially programmers) and deontological principles can help to gain an objective evaluation of machine instructions.	<i>Conceptualization</i> The study provides a proposition of a deontological approach in the context of machine/AI ethics with the aim of avoiding weaknesses of an intuitive system. Based on this, it builds a computational framework.
Horváth (2022) “ <i>AI in Interpreting: Ethical Considerations</i> ” Across Languages and Cultures	<i>HCAI as a moral obligation</i> This work is about human responsibility for developing intelligent systems based on fundamental human values and principles to support the well-being of people. It further considers the moral obligations and duties of individuals based on the question of what is morally right or wrong.	<i>Normative ethics</i> The study mentions the European General Data Protection Regulation (GDPR). Thus, contractarianism is implicitly addressed. <i>Meta-ethics</i> The study provides general thoughts on AI, AI ethics, and ethical principles (data bias and data quality, data privacy and data ownership, transparency).	<i>Enhance HMA with AMA</i> The author argues that AI should help humans reduce cognitive load and facilitate their work. AI ethics should be addressed within the design and not afterthought.	<i>Position paper</i> The author concludes that it is easier to describe rather than to define AI. Several AI ethics issues arise, including data bias/ownership/privacy/quality/transparency.
Jantunen et al. (2021) “ <i>Building a Maturity Model for Developing Ethically Aligned AI Systems</i> ” IRIS Proceedings	<i>Guidelines for ethics communication</i> The study highlights that guidelines are often used to communicate ethical intentions, whereby no single standard has emerged.	<i>Normative ethics:</i> The study argues that AI ethics guidelines should be used as a base when developing and designing ethical AI systems (deontology; implicitly addressed).	<i>HMA as baseline</i> The authors argue that AI ethics guidelines serve as a base for ethical AI systems.	<i>Syntheses</i> The paper highlights that in recent years, numerous guidelines on ethical AI have been published. Due to the high number of different guidelines, standardizations are requested to bridge the gap between research and practice.
Jobin et al. (2019) “ <i>The Global Landscape of AI Ethics Guidelines</i> ” Nature Machine Intelligence	<i>Landscape of ethical principles</i> The authors describe the five most often applied ethical principles that appear in the landscape of AI ethics, namely 1) fairness/justice, 2) non-maleficence, 3) responsibility, 4) privacy, and 5) transparency. The study highlights that companies, public sector organizations, and research institutions have published several guidelines and principles in recent years; however, there is divergence in the interpretation and implementation of these principles.	<i>Normative ethics</i> The study introduces and analyses numerous guidelines as a base for ethical AI, which can be linked to deontology. <i>Applied ethics</i> The study structures existing ethical AI guidelines.	<i>HMA as baseline</i> The authors argue that AI ethical guidelines created by humans are the base for practitioners to achieve ethical AI applications.	<i>Syntheses</i> The work structures and organizes ethical guidelines, principles, and norms of several institutions.
Katznelson & Gerke (2021) “ <i>The Need for Health AI Ethics in Medical School Education</i> ” Advances in Health Science Education	<i>Educate AI ethics for healthcare</i> This work analyzes and applies AI ethics education in the healthcare sector, highlighting that educating ethics from a philosophical point of view is essentially concerned with the question of what is right or wrong.	<i>Applied, normative, and meta-ethics</i> The authors introduce deontology, virtue ethics, and consequentialism, and its practical implementation in a real context. They further elaborate on the question of the nature of (un)rightness.	<i>Training HM:</i> The authors argue that medical students should be trained to develop, implement, engage with, and evaluate AI in a meaningful and ethical way.	<i>Conceptualization</i> The work highlights that there is no standard in medical education of health AI. Six key ethical issues are discussed within health AI: allocation, bias, informed consent, safety, patient

				privacy, and transparency.
Kenny et al. (2021) “Ethics and Standards in the Use of Artificial Intelligence in Medicine on Behalf of the Royal Australian and New Zealand College of Radiologists” Journal of Medical Imaging and Radiation Oncology	<i>From AI ethics to medical ethics</i> This work criticizes AI ethics as being mainly determined by software vendors, whereas the medical context was not considered. Medical ethics covers aspects such as 1) beneficence, 2) beneficence (acting in the best interest of a human being), 3) equity/justice, 4) no harm/non-maleficence, and 5) respect for other human being’s autonomy.	<i>Normative ethics</i> The study analyzes ethics and standards in the context of using AI in medicine. Thus, deontology is implicitly addressed.	<i>Training HMA</i> The authors argue that standards and principles alone are not enough, as people need to be upskilled.	<i>Synthesis</i> The study describes two ethical works released by the Royal Australian and New Zealand College of Radiologists (RANZCR): The ethical principles of AI in medicine and the standards of practice for AI in clinical radiology.
Khan et al. (2022) “AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers” IEEE Proceedings	<i>Applying AI ethics</i> AI ethics are part of applied ethics in the context of ethical issues relating to the development and use of AI systems. Accountability, privacy, and transparency are claimed as the most critical AI ethics principles, whereby lacking monitoring bodies, lack of ethical knowledge, and no legal frameworks are found to be most common AI ethics challenges.	<i>Applied ethics</i> In the study, several AI practitioners and lawmakers were invited to complete an online questionnaire.	<i>HMA as baseline</i> The authors highlight that standards are solely developed by human beings, providing the baseline for ethical conduct.	<i>Quantitative online survey</i> The study asks practitioners and lawmakers regarding AI ethics principles and common AI ethics challenges.
Kieslich et al. (2022) “Artificial Intelligence Ethics by Design. Evaluating Public Perception on the Importance of Ethical Design Principles of Artificial Intelligence” Big Data & Society	<i>Perceptions of ethical AI</i> The study outlines that there is general agreement that ethical AI is important and necessary, but not regarding what AI ethics should look like and how it should be organized. Ethical AI requires regulatory, legal, and political intervention (not primarily tasks for computer scientists).	<i>Normative ethics</i> The study focuses on ethical design principles, which can be linked to deontology. <i>Descriptive ethics:</i> The study contains a survey of 1,204 people to gain insights into their opinions and perceptions of AI ethical theory.	<i>Using HMA to build AMA:</i> The authors argue that ethical principles are needed to mathematically integrate ethics into the specific algorithm and AI application.	<i>Quantitative online survey</i> A survey with 1,204 German citizens, where most value the alignment with ethical requirements, whereby a huge proportion does not demand the implementation of ethical AI.
Kim et al. (2018) “A Computational Model of Commonsense Moral Decision Making” AIES Proceedings	<i>Moral decision-making models</i> The study discusses a model for moral decision-making in the context of autonomous vehicles.	<i>Descriptive ethics</i> The study uses the hierarchical Bayesian model, whereby moral values are analyzed empirically.	<i>HMA-AMA: Distributed moral responsibility:</i> The authors argue that an AI agent must be able to understand the moral basis for human behavior. Even if these actions are based on morally inappropriate values or careless mistakes.	<i>Conceptualization</i> The authors present a computer model to develop the moral judgment capability of autonomous vehicles based on empirical insights.
Knox (2023) “(Re)politicising data-driven education: from ethical principles to radical participation” Learning, Media and Technology	<i>Ethics is not solely a technical matter</i> The work criticizes that ethics is often seen primarily as a form of public relations rather than as an authentic endeavor and interest in essential ethics and morals. By focusing on ethics as a primarily technical matter, the well-founded discussion of moral and justice-oriented aspects is pushed into the background, and the relevance for discussing these is intensified.	<i>Normative ethics</i> The work notes that the basis of AI ethics is formed by various principles and guidelines (deontology). It further describes that thinking about the good and preferable life are certain moral qualities and skills that a person can cultivate internally (virtue ethics).	<i>HMA: shared responsibility</i> The authors argue that AI ethics is grounded in various principles and guidelines, and political achievements need to be made to make use of these guidelines and discuss their content.	<i>Position paper</i> The work discusses the (re)politicization of ethical issues of data-driven technologies.

Koebis et al. (2021) “Bad Machines Corrupt Good Morals” Nature Human Behavior	<i>Influence of AI agents on human’s ethical behavior</i> The study concerns whether AI can have a strong, corrupting influence on people’s ethical behavior. Ethical label development and citizens’ preferences mainly deal with the “direct unethical consequences of AI behavior” (p. 17).	<i>Normative ethics</i> The work implicitly addresses deontology and virtue ethics as the decision to break or follow ethical rules often takes into account how other people would assess the normative standards in this situation.	<i>AMA impacts HMA</i> The authors outline that if AI agents act as role models or advisors, their corrupting power may not exceed those of humans. However, when the AI agent is perceived as a partner, people can avoid unethical behavior.	<i>Conceptualization</i> The paper investigates the influential power of AI on the ethical behavior of humans in a healthcare context.
Koniakou (2023) “From the ‘Rush to Ethics’ to the ‘Race for Governance’ in Artificial Intelligence” Information Systems Frontiers	<i>Shift from ethics to governance</i> This work compares ethical principles with AI governance rules and practices. Thereby, the authors highlight a need to extend human rights obligations in the context of AI governance.	<i>Normative ethics</i> This work explicitly addressed deontology linked to AI governance dealing with regulatory and normative aspects, debates and theories regarding the way of controlling disruptive technologies.	<i>HMA: moral responsibility</i> The authors noted that even though AI systems act autonomic (e.g., algorithmically informed court decisions), regulations and normative theorizations are human-made. Thus, history of AI governance requires further development.	<i>Syntheses</i> The paper provides an overview of the developments within the field of AI governance.
Kornilaev (2021) “Kant’s Doctrine of Education and the Problem of Artificial Intelligence” Journal of Philosophy of Education	<i>Kant in AI curricula</i> This study investigates the impact of AI and ethics in education from an anthropological view. A special focus lies on the influence of whether Kant’s position has to be re-evaluated or not.	<i>Applied ethics</i> The work deals with Kant’s pedagogy.	<i>HMA defines AMA</i> AI challenges the thesis that education is only a human invention. The interpretation depends on whether AI is only a tool or a solely thinking subject.	<i>Philosophical essay</i> This work is a philosophic essay on Kant’s position of “the human being” in the context of AI.
Kowch (2019) “Ethics to Prepare Education for Professional Service Robots in Classrooms” IJCIME Proceedings	<i>Ethics vs. morality</i> This study argues that ethics “is usually associated with our conduct and relationship with others, while morality concerns our conscience and self-relationships” (p. 479).	<i>Applied ethics</i> The study discusses several principles for the use of service robots in classrooms.	<i>Strengthening HMA</i> The authors argue that most professional service robots are either semi- or fully autonomous and feature some form of mobility, thus bringing ethics into service robots getting more and more complex. Educators have to reduce complexity and focus on social issues.	<i>Syntheses</i> The study presents ethical principles in the context of professional service robots in the classroom.
Kučera (2019) “Artificial Intelligence in the Context of Managerial Responsibility” IMES Proceedings	<i>Dilemmas of managerial AI</i> The study identifies ethical dilemmas, such as the dual-use dilemma or risks for democratic stability, when applying AI in the context of managerial responsibility.	<i>Normative ethics</i> This work introduces the deontological school of managerial responsibility for AI, including explicit rules and principles.	<i>Replacing HMA through AMA</i> The paper focuses on the idea of robotization as a replacement of a man by a machine and thus on the (social and ethical) tension between humans and AI.	<i>Philosophical essay</i> The study describes philosophical ideas on the ethical dilemma regarding the application and promotion of AI is addressed.
Lara (2021) “Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates?” Science and Engineering Ethics	<i>Machine ethics</i> This work aims to configure autonomous and robotic systems that can function and make decisions independently according to ethical rules and principles in morally critical situations.	<i>Applied ethics</i> The study concerns whether technological advances should be used to optimize and improve people as moral agents. <i>Normative ethics:</i> The study argues that instead of having AI that makes moral decisions, it should rather nudge individuals to rethink their behavior and make decisions (virtue ethics; implicitly addressed).	<i>Train HMA with AI</i> The authors argue that AI technology helps people to reflectively decide on their own rather than being directly influenced; AI technology is an effective way to train people on moral issues.	<i>Conceptualization</i> This work addresses practical guidelines for moral enhancement through virtual assistants. It is concluded that virtual assistants are better at teaching moral decision-making than human beings.
Li (2021) “Problems with ‘Friendly AI’” Ethics and Information Technology	<i>AI mimicking friendship</i> This work argues that AI ethics often deals with issues concerning the implementation of ethical frameworks. However, ethical issues arise with the perceptions of humans that arise when interacting with	<i>Applied ethics</i> The study discusses how AI should be designed and developed as a friendly AI, including, for example, mimic aspects.	<i>AI impact HMA</i> The authors argue that a friendly AI would be able to assist humans with ethically and morally correct actions. Currently, however, AI is limited in its capabilities and lack of awareness.	<i>Philosophical essay</i> The work deals with friendly AI, which includes mimicking issues of human friendship.

	human-like friendly AI, mimicking friendship.			
Lim & Kwon (2021) “A Study on the Modeling of Major Factors for the Principles of AI Ethics” DG.O Proceedings	<i>Standardization of AI principles</i> This study investigates major factors of the different ethical principles of AI ethics that exist across several countries to identify some standards.	<i>Normative ethics:</i> The study presents principles on what is morally, thus addressing deontology implicitly. <i>Applied ethics</i> The study analyzes the ethical principles of 15 different countries.	<i>HMA: shared responsibilities</i> The authors illustrate that AI ethic principles are presented by different governmental institutions.	<i>Conceptualization</i> The study presents a model covering 12 major factors of different principles of AI ethics across 15 countries.
Lloyd (1985) “Frankenstein’s Children: Artificial Intelligence and Human Value” Metaphilosophy	<i>Ethical issues of human-like AI</i> This work provides moral considerations on ethical issues that arise when AI increasingly performs human-like cognitive processes.	<i>Normative ethics</i> The work argues that “Consciousness is the Foundation of Intrinsic Value: Each of us is a locus of intrinsic value. But on what does this fundamental principle rest? I will suggest here that it is because we are conscious that each of us has value as a Kantian end-in-itself” (p. 312). Thus, it addresses virtue ethics implicitly and deontology explicitly.	<i>HMA guiding AMA</i> The author argues that ethical AI should be determined by human values as AI lacks consciousness.	<i>Position paper</i> The paper discusses ethical issues that appear when developing cognitively stimulating AI.
Mabaso (2021) “Computationally Rational Agents can be Moral Agents” Ethics and Information Technology	<i>Robots as moral agents</i> This work is concerned with the development of robots and other systems that are capable of making ethical and moral decisions.	<i>Meta-ethics</i> This work describes computational and philosophical facets of artificial moral agency.	<i>Rational agency</i> The authors argue that “Moral agency implies a certain understanding and knowledge of what is good and what is bad (morality) and being able to discern what is right from what is wrong (ethics)” (p. 139).	<i>Philosophical essay</i> The work provides a combination of a philosophical and scientific perspective on artificial moral agency.
Manna & Nath (2021) “Kantian Moral Agency and the Ethics of Artificial Intelligence” Problemos	<i>Kantian moral agency</i> The study argues that when machines have moral agency, their “duty” should be performed with “freedom of will” and happiness” referring to Kant. However, the “freedom of will” is a sacrosanct concept that does not fit to “deterministic models of agency” (p. 139).	<i>Meta-ethics</i> The work discusses whether AI agents can have a Kantian moral agency.	<i>AI simulates HMA</i> The authors argue that “AI robots can be the perfect simulacrum of human moral agents without any conscious moral agency, and their rationality is algorithmic” (p. 141). Thus, the moral behavior of robots depends on the human programmers, which is why the moral agency of robots differs from the genuine Kantian agency. In some situations, robots may perform better than humans, but their behavior is amoral.	<i>Philosophical essay</i> The work elaborates on Kantian moral agency in the AI context.
Mäntymäki et al. (2023) “Designing an AI Governance Framework: From Research-based Premises to Meta-requirements” ECIS Proceedings	<i>Epistemic AI ethics</i> The work addresses concerns that decisions are made on the basis of incorrect or inconclusive findings.	<i>Normative ethics</i> The work implicitly addresses deontology by stating that “AI governance comprises tools, rules, processes, procedures, and values that aim to ensure the legally compliant and ethically aligned development and use of AI” (p. 2).	<i>HMA: shared responsibilities</i> The authors argue that human-centric AI has been the focus, leading to several AI ethics principles.	<i>Qualitative content analysis</i> The work provides practical contributions by seven meta-requirements for the design of AI systems as a starting point for AI auditing frameworks to bridge the gap between AI ethics principles and practice.
Martin (2019) “Ethical Implications and Accountability of Algorithms”	<i>Communicating ethical implications</i> Information on algorithm accountabilities during development and	<i>Normative ethics</i> The work implicitly addresses consequentialism by stating, “In that algorithm create moral	<i>AMA-HMA: Distributed moral responsibility</i> The authors criticized that algorithm’s structure impacts the lives of human beings,	<i>Position paper</i> This work focuses on the opinion that algorithmic has a huge impact on human beings,

Journal of Business Ethics	implementation processes is crucial.	consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity” (p. 835).	whereby the responsibility is not finally defined.	covering access to rights and social goods.
Martinho et al. (2021) “Computer Says I Don’t Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence” Minds & Machines	<i>Machine ethics</i> This work argues that machine ethics is controversially discussed and deals with ethical principles as well as moral decision-making in the development of machines to ensure that the machines’ behavior towards human users is ethically appropriate.	<i>Descriptive ethics</i> The study investigates moral uncertainties empirically.	<i>Rising AMA:</i> The authors argue that AI systems are increasingly acting autonomously, leading to greater (moral) agency.	<i>Quantitative online survey</i> This work evaluates a model (validation/testing) as an empirical approach to investigate moral uncertainty in the AI-context.
Mayer et al. (2021) “How Corporations Encourage the Implementation of AI Ethics” ECIS Proceedings	<i>Human-centered AI implementation</i> The study implies that the user plays a significant role in fostering morality in human-AI interaction.	<i>Applied ethics</i> The study focuses on how companies can encourage the implementation of AI ethics. <i>Normative ethics</i> The study focuses on ethical guidelines (deontology; implicitly addressed) and the user, who needs to foster ethics in human-AI interaction, which is linked to virtue ethics (implicitly)	<i>HMA guides AI</i> The authors argue that human values should guide AI’s self-learning and autonomous operations.	<i>Interviews</i> The study uses nine semi-structured expert interviews to identify measures and mechanisms to encourage employees for ethical AI practices.
McDonald & Pan (2020) “Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact” PACM Proceedings	<i>Student’s perceptions on AI and ethics</i> This work focuses on “how to prepare future generations to recognize and grapple with the ethical concerns of a range of issues plaguing AI” (p. 147:10).	<i>Normative ethics</i> This work focuses on highlighting the importance of virtue ethics (implicitly) in human-AI interaction (e.g., “a ‘debiased’ predictive algorithm that forecasts where crime will occur if used by law enforcement to harass and oppress communities of color, becomes ‘unfair’ by virtue of the use to which it is put” (p. 147:4))	<i>HMA-AMA: shared/distributed moral responsibility</i> The authors argue that human moral agency may have certain limits since “humans may be poor moral arbiters, not suitably wired to solve difficult ethical dilemmas.” (p. 147:3) AI agency is somehow difficult to control “it may be unreasonable to expect designers to fully anticipate the behaviors of code that have the potential to suppress freedom” (p. 147:3).	<i>Interviews</i> The paper conducts twenty qualitative interviews with graduate students to understand their ethical considerations about AI implications linked to how AI designs can be biased.
Millar (2016) “An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars” Applied Artificial Intelligence	<i>Hybrid ethics</i> The study argues that ethical decision-making in robots requires a hybrid approach that combines stakeholders’ different moral interests and requires human inputs. Yet, AI is seen as an active participant in the user-robot relationship.	<i>Applied ethics</i> The paper investigates how ethics can be integrated into the design to “design a tool that can be used to avoid creating robots that subject users to morally questionable paternalism” (p. 793).	<i>Distributed moral responsibility: explicit ethical agents</i> The authors suggest that ethical considerations need to see AI as active participants and ethical decision-makers, yet “humans need to implement their morality to AI as AI has no moral by itself” (p. 792).	<i>Synthesis</i> The main objective of the paper is a design tool aimed at preventing the development of robots that impose morally questionable paternalism on users.
Minkinen & Mäntymäki (2023) “The Institutional Logics Underpinning Organizational AI Governance Practices” SCIS Proceedings	<i>AI governance</i> In the study, AI ethics is defined through AI governance practices in the form of ethical requirements for AI (like fairness, transparency, etc.), and humans should be informed of them through feedback.	<i>Normative ethics</i> The study utilizes AI governance literature and refers to literature on institutional logic and institutional pluralism and relates it to organizations’ AI governance practices, which can be linked to contractarianism.	<i>AMA: shared responsibility</i> The paper focuses on how to translate AI governance and ethical principles to AI so that AI acts according to human values.	<i>Synthesis</i> The paper summarizes AI governance practices from literature and the presentation of four institutional logics on how to conceptualize the relationship between organizations’ AI

				governance and AI governance practices.
Morley et al. (2021) “ <i>Ethics as a Service: A Pragmatic Operationalisation of AI Ethics</i> ” Minds & Machines	<i>Operationalizing ethics</i> The study describes AI ethics principles and rules, which should be understood as “procedural regularity” (p. 247) and not mainly principle-based. The authors introduce “ethics as a service” (p. 239).	<i>Applied ethics</i> The study aims to make AI ethics useful for practitioners by referring to a concept that the authors call “ethics as a service.”	<i>Hybrid moral agency:</i> The authors suggest highly complex systems where AI also has agency by stating, “In complex systems, agents interact with each other and with other systems in unexpected ways, making their response to change unpredictable and non-linear.” (p. 253)	<i>Synthesis</i> By relying on prior literature, the paper aims to make AI ethics useful for practitioners by avoiding purely principle regularity and heading toward procedural regularity.
Munoko et al. (2020) “ <i>The Ethical Implications of Using Artificial Intelligence in Auditing</i> ” Journal of Business Ethics	<i>Humans in control</i> This study argues that human-AI ethics requires humans to control AI, since AI is limited concerning intuitive intelligence and empathy. The authors state, „In the ideal situation, the auditor would be aware of the type of AI artifact in use ... and what its benefits and risks are. The auditor would not relinquish their responsibility to the technology ... they would perceive the technology as a complement to their professional judgment and exhibit an appropriate level of skepticism” (p. 5).	<i>Applied ethics</i> The paper focuses on ethical issues arising in practice, while the paper relies on ethical principles and codes of conduct.	<i>Hybrid moral agency: implicit ethical agents</i> The authors argue that AI has a certain level of autonomy and self-learning ability by stating that “AI systems “learn” from data and can self-evolve over time due to exposure to new data, without explicitly being programmed by a human” (p. 2). However, “AI might be limited when it comes to complex judgments that require professional skepticism” (p. 4).	<i>Synthesis & case study</i> Based on prior studies and reported use of technology by auditing firms, the paper investigates which ethical issues may arise in auditing when implementing AI tools in their work practices.
Neubert & Montañez (2020) “ <i>Virtue as a Framework for the Design and use of Artificial Intelligence</i> ” Business Horizon	<i>Relevance of virtue ethics</i> This study highlights that virtue ethics performed by humans is needed to guarantee ethical human-AI interaction.	<i>Applied ethics</i> The study highlights real-world scenarios in which they transfer their views on virtue ethics. <i>Normative ethics</i> The study focuses on how to foster good behavior in developing and using AI (linked to virtue ethics) as frameworks and as principles that may not always work in practice (linked to deontology).	<i>HMA: humans remain responsible</i> The paper clearly focuses on virtue ethics and suggests that humans have moral agency and that virtue is crucial to guarantee responsible behavior in human-AI interaction.	<i>Quantitative text analysis</i> The authors encounter ethical challenges on Google and highlight the importance of virtue ethics for the development and usage of AI. They suggest a framework for ethical decision-making.
Nevanperä et al. (2021) “ <i>Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers!</i> ” KES Proceedings	<i>Mixing ethical views</i> The study argues that human-AI ethics should be fostered through a mix of duty and virtue. There should be ethics by design, and AI should be developed according to human values, referring to values-in-design.	<i>Applied ethics</i> The study outlines ethical guidelines for AI developers that are designed according to ethical principles. <i>Normative ethics</i> The study compares and combines virtue ethics and deontology.	<i>AMA: explicit ethical agents</i> The authors perceive AI as having limited moral agency and humans remain responsible. They state, “Even though machine ethics consider artificial intelligence at least partially as a moral agent, in the real life we cannot at least not yet hold artificial intelligence and its algorithms accountable for the decisions it makes” (p. 2333).	<i>Synthesis</i> By building on prior research and ethical guidelines, this paper suggests a process for designing ethical guidelines according to design science research for AI-based technologies that support the life of older people.
Niederman & Baker (2023) “ <i>Ethics and AI Issues: Old Container with New Wine?</i> ” Information Systems Frontier	<i>Using the Turing test for the development of ethical AI</i> To foster ethics in human-AI interaction, the authors suggest relying on a combination of developed solid Turing tests and guidelines and fostering “broad maxims stating ideal relationships of AI robot to human across a wide range of circumstances and contingencies” (p. 21).	<i>Normative ethics</i> The paper investigates how to address ethical issues in guidelines and reflections on whether machines should “continue embodying increasingly observable agency by demonstrably exercising consciousness, benevolence, and/ or imagination” which can	<i>Hybrid moral agency: implicit ethical agents</i> The authors’ thoughts can be linked to the idea that AI has limited moral agency as they state, “We consider the escalation of the distancing of human agency from the effects of mechanical actions. There is a difference between direct effects of purposeful human activity and indirect effects of machinery that responds to unintended	<i>Synthesis</i> The paper investigates prior research and aims to categorize AI ethical issues and investigate which issues may emerge from AI use that need to be addressed in ethical guidelines.

		be linked to different ethical views (p. 10).	actions or neglectful inattention rather than explicit, purposeful agency” p. 10).	
Obi & Gray (2023) “Auditing Practitioner Judgment for Algorithmic Fairness Implications” IEEE Proceedings	<i>Judging algorithmic fairness</i> The paper emphasizes the key role that designers play in their judgment decisions when designing AI.	<i>Descriptive ethics</i> The study reflects on the role of practitioners in developing ethical AI systems and thereby connects practitioner judgments to related fairness and justice implications.	<i>HMA: shared responsibility</i> The paper shows the key role that designers play in developing AI.	<i>Case Study</i> Based on exploratory observations, the paper aims to investigate the sensemaking of software engineers that influences their decisions when developing AI and how to foster fairness through ethical guidance, support, and interventions.
Ong (2021) “An Ethical Framework for Guiding the Development of Affectively Aware Artificial Intelligence” ACII Proceedings	<i>Different views on ethics</i> The paper suggests that different stakeholder’s views and regulations are relevant for ethical human-AI interaction.	<i>Normative ethics</i> The paper stresses the importance of ethical guidelines (deontology) and the moral values of developers and users in assigning responsibility for ethical issues (virtue ethics). By arguing that AI ethics should be based on what is best for society, consequentialism is further implicitly addressed.	<i>HMA: shared responsibility</i> The author outlines that AI agency is limited as it does not consider context compared to humans. Thus, different stakeholders should peer-review the technologies developed as AI is “fundamentally limited in this endeavor of reading internal states solely from external behavior” (p.3).	<i>Synthesis</i> The authors suggest a framework that highlights the implications and consequences of affective-aware AI based on prior research.
Palmer & Schwan (2024) “More Process, Less Principles: The Ethics of Deploying AI and Robotics in Medicine” Cambridge Quarterly of Healthcare Ethics	<i>Process over principles</i> The study highlights that ethics ought to emphasize process over principles; policy outcomes should be influenced by diverse stakeholders.	<i>Applied ethics</i> The study reflects on ethical considerations in practice linked to virtue ethics by stating, “Technology must be sensitive to the unique normative and teleological features of this practice” (p. 2).	<i>HMA: moral intention</i> The paper focuses on the ethical employment of AI as a moral agency rather than with humans.	<i>Position Paper</i> The paper suggests a practical model (called participatory deliberative conservatism) that helps medical institutions make ethical deployment decisions about advanced technologies (such as AI). The authors suggest “more process, less principles.”
Pansoni et al. (2023) “Artificial Intelligence and Cultural Heritage: Design and Assessment of an Ethical Framework” CIPA Proceedings	<i>Need for AI governance</i> The study highlights that there is as need for ethical guidelines guiding AI by stating that “there is a need to establish an ethical framework to morally regulate the use of AI” (p. 1152).	<i>Normative ethics</i> The study’s main focus lies on principles for establishing a moral framework, which is linked to deontology.	<i>AMA: responsibility attribution</i> The authors argue that algorithms may have certain degrees of agency, though not necessarily a “moral” one, by stating, “where duty falls when algorithms have varied degrees of agency to act is another problem that arises with the application of AI” (p. 1152)	<i>Synthesis</i> Based on prior literature and guidelines for the trustworthy implementation of AI, the paper suggests a framework that consists of six core principles to identify ethical risks.
Parlangeli et al. (2020) “How Guilty is a Robot who Kills Other Robots?” IEEE Proceedings	<i>Human perceptions on responsibilities</i> The paper investigates humans’ perceptions of responsibility for unethical outcomes.	<i>Descriptive ethics</i> The paper states that assigning moral agency to AI and taking responsibility depends on individual factors and personality.	<i>HMA-AMA: moral intention</i> The authors address the fact that humans assign moral agency to AI when making moral judgements.	<i>Experiment</i> This is a study based on 293 students making moral judgments on a story to investigate how the type of victims and personality traits affect moral judgment.
Polyviou & Zamani (2022)	<i>Human perceptions of AI strategies</i>	<i>Descriptive ethics</i> The study investigates different stakeholder	<i>Protecting HMA</i> The authors perceive AI algorithms as black boxes,	<i>Qualitative content analysis</i>

“Towards Europe’s AI Strategy: Desires & Realities” ECIS Proceedings	This work refers to individuals’ perceptions of Europe’s AI strategy.	perceptions of AI frameworks and strategies.	suggesting that humans need to keep their autonomy in human-AI interaction.	This study explores stakeholders’ perceptions about the forthcoming European AI strategy, which reveals six dimensions that are considered crucial for a responsible and ethical AI strategy.
Pruski (2023) “Ethics Framework for Predictive Clinical AI Model Updating” Ethics and Information Technology	<i>Process over principles</i> This work describes ethics as rather principle-based, and AI ethics need to be updated according to societal values. This refers to rather a process view.	<i>Applied ethics</i> The paper suggests a framework to show best practices on how to choose the ethical strategy to obtain the information required to update clinical predictive AI models by stating, “as the ethnic make-up of societies changes it needs to be ensured that the population on which the AI was trained reflects its current composition” (p. 2).	<i>AMA: explicit agents that require human oversight</i> The study defines AI agency through code and data “if the AI utilises a black box algorithm [it is hardly] feasible to assess the factors which the AI utilises” (p. 11). Thus, AI algorithms need to be updated by humans.	<i>Synthesis</i> The paper suggests a framework to show best practices on how to choose ethical strategies to update clinical predictive AI models.
Raji et al. (2021) “‘You Can’t Sit With Us’: Exclusionary Pedagogy in AI Ethics Education” FACCT Proceedings	<i>Interdisciplinary AI curricula</i> This work argues that ethics should have interdisciplinary recognition by stating that “challenges are inherently interdisciplinary, requiring the cooperation of stakeholders of varying expertise in business, law, and other domains in order to meaningfully address in the real world” (p. 515).	<i>Meta-ethics</i> The study challenges previous assumptions on norms in education by stating, “in order to address the social impact of technical systems, including AI, we need to revisit the way we think about the norms of AI ethics education” (p. 522).	<i>HMA: shared responsibility</i> The study highlights that “there is an underlying tension between individual agency and bureaucratic control with actual responsibility lying somewhere in between” (p. 517).	<i>Position Paper</i> Through analysis of computer science education literature and a review of college-level course in AI ethics, the paper proposes a shift towards a substantively collaborative, holistic, and ethically generative pedagogy in AI education.
Rismani & Moon (2023) “What Does it Mean to be a Responsible AI Practitioner: An Ontology of Roles and Skills” AIES Proceedings	<i>Skills and roles for responsibilities</i> This work argues that AI in education requires humans who are skilled and have the competencies to behave in a responsible way (besides their given roles making them responsible). Thus, interdisciplinary teams are required.	<i>Normative ethics</i> The study suggests skills, qualifications, etc., that are relevant to ethical practitioners’ sense of responsibility, which can be linked to virtue ethics.	<i>HMA: shared responsibilities</i> The authors focus on human moral agency by suggesting a competency framework splitting responsibilities among AI practitioners based on roles and competencies.	<i>Interviews</i> Based on an analysis of job postings and interviews, the paper suggests a competency framework by outlining the roles and responsibilities of AI practitioners and what skills and qualifications they need to conduct their roles.
Roberts & Montoya (2022) “Contextualizing Artificially Intelligent Morality: A Meta-Ethnography of Theoretical, Political and Applied Ethics” FICC Proceedings	<i>Different perspectives on AI ethics require hybrid approaches</i> The study highlights that AI ethics need to be viewed from different perspectives by stating, “ethics around AI is complex, and especially when deployed globally, its implementation needs to be considered from multiple angles” p. 2).	<i>Meta-ethics</i> The study follows holistic considerations of AI morality (top down and bottom up) from a theoretical, technical, and political stance.	<i>HMA: no ethical agents</i> The authors suggest that “machines cannot be held responsible for what they learn from people, ever. Machines do not have an inherent conscience or morality as humans do” and “the problems that arise in AI often reflect the problems we have in society” (p. 19).	<i>Philosophical essay</i> The paper seeks to increase the understanding on ethics of AI and how it can be implemented in AI by analyzing it from a political, technical, and theoretical perspective.
Roberts et al. (2023) “Governing Artificial Intelligence in China and the	<i>AI governance: comparing aims with moral values</i> This work argues that human AI ethics depends on context, culture, and society and should be seen as ethical pluralism guided by policy makers.	<i>Normative ethics</i> The study argues that moral norms come from mutual agreements (linked to contractarianism), and their interpretation can	<i>HMA: shared responsibility</i> The authors suggest to “establish an AI accountability mechanism to clarify the responsibilities of developers, users, beneficiaries, etc.” (p. 87).	<i>Synthesis</i> By comparing China and the EU, the paper outlines policy suggestions for policymakers to promote ethical

European Union: Comparing aims and Promoting Ethical Outcomes” The Information Society		differ depending on the context. The authors state, „we adopt a stance of ethical pluralism that recognizes agreement might be limited to the existence and nature of fundamental values, but that their interpretation may differ depending on the context” (p. 80).		outcomes of AI strategies.
Rodgers et al. (2023) “An Artificial Intelligence Algorithmic Approach to Ethical Decision-Making in Human Resource Management Processes” Human Resource Management Review	<i>Moral reasoning in HRM</i> This work focuses on human reasoning in ethical decision-making in HRM processes.	<i>Applied ethics</i> The study provides ethical considerations in decision-making in HRM. <i>Normative ethics</i> The study combines different normative views, such as ethical egoism, deontology, utilitarianism, relativism, virtue ethics, and ethics of care, in the HRM context.	<i>AMA: implicit ethical agents</i> This study considers AI to have a certain degree of morality by stating, “AI algorithms can incorporate ethical considerations, decision-making processes, and managers’ knowledge to determine the most appropriate HRM strategy in each situation” (p. 2).	<i>Position paper</i> The paper suggests a model with six dominant algorithmic pathways to provide possible solutions to foster ethical decision-making in HRM.
Saariluoma & Leikas (2020) “Designing Ethical AI in the Shadow of Hume’s Guillotine” IHSI Proceedings	<i>Ethical AI</i> This study reflects on moral values by stating that “facts are true or false and epistemic in nature, and in contrast, values are either more or less justified, and thus are ethical or moral” (p. 595).	<i>Applied ethics & normative ethics</i> The paper raises questions about whether morality can be integrated into AI by reflecting on different ethical views.	<i>AMA: explicit ethical agents</i> In the paper, AI is suggested to have ethical rules but no ethical norms since one cannot derive values from facts by stating, “it is easy to provide ethical rules for intelligent machines. ... but there are no principled obstacles to programming ethical norms into machines” (p. 597).	<i>Philosophical essay</i> The paper is questioning to which extent AI can be assigned moral values.
Schlimbach & Khosrawi-Rad (2022) “Towards Ethical Design Features for Pedagogical Conversational Agents” AMCIS Proceedings	<i>Pedagogical conversational agents</i> This study aims to develop requirements and specific features for pedagogical conversational agents that uphold ethical standards.	<i>Applied ethics</i> The study is concerned with the ethical design of pedagogical conversational agents.	<i>AMA: explicit ethical agents</i> AI morality is influenced by data and humans, and AI is limited in its agency. Humans are accountable for AI output.	<i>Design science approach</i> Based on a co-creation process embedded in design science research with a total of 40 students, the paper derives requirements and concrete features for ethically responsible pedagogical conversational agents and reflects them along with the OECD principles for trustworthy AI.
Schneider et al. (2023) “How Individuals can Shape AI Through Data - An AI Literacy and Morality Perspective” ECIS Proceedings	<i>Individuals’ literacy and morality shape AI design</i> This study “focuses on data donors, who are common users of AI systems, wondering whether to provide (or modify) data to train an AI system to achieve a specific goal such as improving or worsening overall system performance” (p. 10).	<i>Normative ethics</i> Analyze peoples’ ethical considerations when donating data and “address these concerns by using Kant’s perspective” (p. 5). Thus, deontology and virtue ethics are addressed.	<i>HMA: moral intention</i> AI morality is influenced by data and humans, and AI is limited in its agency. Humans can foster morality in AI but may also manipulate it	<i>Design science research & case studies</i> The paper provides a conceptual understanding of the impact of donating data of different quality on the AI system’s performance and provides insights into the ethical considerations of data donors.
Seo & Thorson (2022) “Computation,	<i>Aligning rules with context</i> This study focuses on AI alignment and suggests that AI	<i>Descriptive ethics</i> The paper introduces rules, including moral	<i>HMA: limited human agency</i> The authors argue that human ethical and moral	<i>Philosophical essay</i> The paper proposes a framework

<i>Rule Following, and Ethics in AIs</i> ” HICSS Proceedings	ethics is rule-based and highly depends on context (“ethics rules, while propositional do not generally take on truth values ... If H partially describes an environment, and C denotes a partial state of affairs in some possible ‘future’ environment, then R can be interpreted as positing a relation between two environments” p. 1077)	and ethics rules, as linguistic entities and illustrates how they can be encoded and treated computationally.	reasoning could be understood as fundamentally computational and does not involve some special human-only force or quality.	centered on rule-following to examine issues in ethics, AI alignment, and computational models.
Seppälä et al. (2021) “From Ethical AI Principles to Governed AI” ICIS Proceedings	<i>Implementing ethical AI principles</i> The study highlights practices on how to implement ethical AI principles and suggests that “both technical and non-technical methods are required for AI principles to be implemented” (p. 2).	<i>Applied ethics</i> The study focuses on the implementation of ethical principles in practice.	<i>HMA: moral intention</i> The authors argue that humans should have control and oversight in human-AI interaction, and individuals should always have the possibility to opt out of human-AI interaction.	<i>Interviews</i> By relying on expert interviews, this paper identified that ethical AI principles are implemented through four sets of practices: i) governance, ii) AI design and development, iii) competence and knowledge development, and iv) stakeholder communication.
Serafimova (2020) “Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement” Humanities & Social Science Communication	<i>AI ethics and machine ethics</i> The paper examines AI ethics within the framework of machine ethics, emphasizing the importance of instilling ethical principles in AI to ensure morally appropriate interactions with human users.	<i>Normative ethics</i> All three major normative ethical theories are considered within the study: deontology, utilitarianism, and virtue ethics, as reflected in the conceptualization of “deontological, utilitarian, and virtue ethical machines” (p. 2).	<i>HMA: full ethical agents</i> The authors argue that although humans are full ethical agents, AI, even when exhibiting moral behavior, does so through human-directed updates. This raises questions regarding the authenticity and limits of AI’s moral perfection.	<i>Philosophical essay</i> The work compares three normative theories to describe moral AI scenarios.
Seymour (2018) “Artificial Intelligence Is No Match for Human Stupidity: Ethical Reflections on Avatars and Agents” ACIS Proceedings	<i>AI perceptions shape ethics in human-AI interaction</i> This study explores how human narratives and perceptions shape ethical considerations in AI, advocating for a more grounded and realistic evaluation of human-AI interactions.	<i>Normative ethic</i> The study implicitly engages with virtue ethics, particularly through the idea that humans project their values onto AI and assign agency based on their notions of fairness and impartiality.	<i>HMA: hybrid agency</i> The author advocates for the development of realistic AI agents by analyzing the “hybrid agency” (p. 9) humans confer upon AI, which is shaped by human ethical values and perceptions of fairness, guiding AI behavior in complex social contexts.	<i>Philosophical essay</i> The work is an essay that strives for the development of realistic AI agents.
Shank et al. (2019) “When are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions after Individual and Joint Decisions” Information, Communication & Society	<i>Moral attribution of responsibility within joint decision-making</i> The study investigates the ethics of human-AI interactions, particularly moral attribution of responsibility in contexts where joint decision-making occurs, and moral attributions are based on the outcomes of these collaborations.	<i>Normative ethics</i> The work implicitly addresses consequentialism, focusing on the ethical implications of AI-driven outcomes and the assignment of responsibility for moral transgressions. <i>Descriptive ethics</i> The work contains an empirical investigation of responsibility attribution and moral judgment.	<i>HMA to AMA: hybrid agency</i> The study reveals that moral attributions in human-AI joint decisions are influenced by the degree of human control and the nature of moral faults, highlighting the complexity of moral responsibility. The study finds that humans overseeing AI are judged less harshly than those who merely receive AI recommendations.	<i>Experiment</i> The study performs an experiment with 453 MTurk workers facing real-world events stimulating joint decision-making among human-AI teams.
Shevlin (2021) “How Could We Know When a Robot was a Moral	<i>AI robots as moral patients</i> This work addresses the critical question of when and under what conditions artificial beings might be granted moral	<i>Meta-ethics</i> This work takes an ontological approach, exploring five strategies to determine whether AI	<i>AMA: hybrid agency</i> The authors argue that to consider AI robots as moral patients requires attributes such as intuition, empathy,	<i>Philosophical essay</i> The study discusses conditions under which AI-based robots might be

<i>Patient?</i> ” Cambridge Quarterly of Healthcare Ethics	status, engaging deeply with the moral patient versus moral agent debate.	can be considered artificial psychological moral patients, focusing on cognitive equivalence with humans.	autonomy, and behavioral equivalence. Hence, moral agency depends on the extent humans can draw equivalences to other beings they interact with.	considered moral patients in the context of healthcare.
Shih et al. (2021) “ <i>Learning Ethics in AI—Teaching Non-Engineering Undergraduates through Situated Learning</i> ” Sustainability	<i>AI ethics in non-technological classes</i> This work focuses on raising the awareness of AI ethics through increased interaction and teamwork in universities as this changes humans' AI understanding and attitudes as antecedents for making humans aware of ethical issues.	<i>Applied ethics</i> The study advocates for a situation-based, hands-on learning approach with AI to heighten awareness of AI ethics among students.	<i>HMA-AMA: moral intentionality</i> The authors argue that increased interaction with AI fosters greater awareness of ethical issues in human-AI interactions. Suggests that such engagement can influence human moral agency without attributing moral agency to the AI itself.	<i>Case study</i> The paper provides a case study on educational AI initiatives in Taiwan with 328 first-year students with a non-technical background.
Shneiderman (2020) “ <i>Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems</i> ” ACM Transactions on Interactive Intelligent Systems	<i>HCAI guidelines</i> This study proposes guidelines for developing AI systems that are reliable, safe, and trustworthy, underpinned by robust development practices, management strategies, and independent certification. Thereby, AI systems should be reliable, safe, and trustworthy through profound development practices (team level), adequate management strategies (organizational level), and certification by independent oversights (industrial level).	<i>Applied ethics</i> The paper integrates various levels of governance (team, organizational, industrial) to ensure adherence to AI principles, emphasizing deontological principles related to industrial and societal standards. <i>Normative ethics</i> The study implicitly refers to contractarianism as industrial and societal standards of AI governance policies are addressed, which can be further related to deontology.	<i>HMA: distributed agency</i> The author calls for a form of shared responsibility of different human-related stakeholders at three different levels of governance (distributed agency): team (AI developers), organization (managers), and industry (regulators).	<i>Conceptualization</i> The paper presents a comprehensive AI governance framework for the development of AI design principles and ethical standards in diverse industrial sectors (e.g., finance and consultancy).
Siapka (2022) “ <i>Towards a Feminist Metaethics of AI</i> ” AIES Proceedings	<i>AI ethics through a feminist meta-ethical lens</i> This study focuses on “moral judgments in a non-normative way” by not solely focusing on what ethics is in human-AI interaction “but also what our engagement with it should be like” (p. 665). Thus, it explores AI ethics through a feminist meta-ethical lens, challenging traditional notions of ethics in human-AI interaction and advocating for a relational approach that considers gender dynamics.	<i>Meta-ethics</i> The paper employs a feminist meta-ethical perspective to interrogate the relational aspects of AI ethics, drawing on insights from philosophers with practical experience.	<i>HMA: moral intentionality</i> The authors critique conventional approaches to AI ethics by incorporating feminist perspectives, arguing that moral ignorance does not absolve humans from responsibility, even when not directly involved in AI interactions. This highlights the broader societal implications of AI ethics from a European feminist viewpoint.	<i>Philosophical essay</i> The work considers AI ethics from a societal and feministic-European perspective.
Siau & Wang (2020) “ <i>Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI</i> ” Journal of Database Management	<i>AI ethics and ethical AI</i> The study highlights the distinction between the ethics of AI (ethical principles guiding AI) and ethical AI (AI behavior that aligns with these principles) in human-AI interactions. Ethical principles (ethics of AI) can lead to ethical behavior of the AI (ethical AI) within the interaction.	<i>Normative ethics</i> The paper examines virtue ethics (think critically about moral values) and deontology (ethics as a set of principles). <i>Applied ethics</i> The paper further refers to computer ethics and machine ethics as an applied ethics branch.	<i>HMA-AMA: hybrid agency</i> The authors develop a framework based on AI ethics objectives and ethical principles from eight institutions, emphasizing the integration of human and machine agency in fulfilling social roles and responsibilities. They propose a distributed ethical agency model that accounts for the interplay between human and machine ethical considerations. as “[r]esponsibility means that machines fulfil some social role that carries with it some assumed responsibilities” (p. 81).	<i>Conceptualization</i> The paper provides a framework based on the review of AI ethics objectives and ethical principles of eight institutions.

Smith & Vickers (2021) “Statistically Responsible Artificial Intelligences” Ethics and Information Technology	<i>Responsible AI</i> The authors argue that “[o]nly through continued human/AI interaction can we determine whether AIs can be moral agents, full or marginal” (p. 493). Hence, ongoing human-AI interaction is crucial for understanding AI’s potential as moral agents, with AI capable of mimicking human behavior but lacking intrinsic moral agency.	<i>Normative ethics</i> The study implicitly engages with virtue ethics through the application of Strawson’s (2008) approach to moral responsibility, rooted in human existence and interaction.	<i>AMA: full ethical agents</i> The authors contend that while AI can imitate responsible behavior, it lacks genuine moral agency. Emphasizes the importance of human interaction in shaping ethical AI, underlining that moral responsibility is inherently a human trait.	<i>Conceptualization</i> The work uses a philosophical framework that builds upon Strawson’s (2008) approach (meaning that moral responsibility as reactive actions are rooted in human existence and, thus, can only occur through interactions).
Stahl et al. (2022) “A European Agency for Artificial Intelligence: Protecting Fundamental Rights and Ethical Values” Computer Law & Security Review	<i>AI governance: The European AI Act</i> This work examines AI ethics within the broader context of technology governance, focusing on the European AI Act as a mechanism for safeguarding fundamental rights and ethical values in AI deployment.	<i>Applied ethics</i> The study integrates human rights theory and European ethics (e.g., the European AI Act) to evaluate the implications of AI governance for fundamental rights, with a focus on responsible AI. <i>Normative ethics</i> The study further implicitly addresses contractarianism referring to AI governance policies.	<i>HMA: distributed agency</i> The authors advocate for the establishment of a European Agency for AI to ensure compliance with the AI Act. The authors highlight the ethical significance of this initiative, emphasizing the need for transparent and accountable AI governance mechanisms.	<i>Interviews</i> The study builds upon 18 expert interviews demonstrating that different regions form different regulations and regulatory agencies.
Stahl et al. (2023) “Exploring Ethics and Human Rights in Artificial Intelligence – A Delphi Study” Technological Forecasting & Social Change	<i>Limitation of AI ethics principles</i> This study argues that while most of the AI ethics research is based on principles (mostly driven by the biomedical ethics community), research also has shown that principles are somehow limited. Therefore, ethics in human-AI interaction is way more complex and due to the lack of agreement on responsibilities, more than principle-based ethics is required.	<i>Normative ethics</i> The work refers to all well-known ethical theories of normative ethics by stating, “Prominent ethical theories point to the importance of consequences of actions in judging their ethical quality (Bentham, 1789; Mill, 1861); others underline the importance of duty and logic (Kant, 1797, 1788), whereas others focus on personal character and virtue (Aristotle, 2007)” (p. 2).	<i>AMA-HMA: full ethical agents</i> The authors follow the definition of McCarthy et al. (2006) for AI by claiming that when AI can simulate every aspect of human intelligence (HI), it is hard to set clear boundaries between AI and HI and their responsibilities (as both seem to be full ethical agents). Also, within their Delphi study participants demonstrated a lack of agreement on responsibility attribution (and thus, unaligned human moral agency) for developing, promoting, deploying, and using AI.	<i>Delphi study</i> The work uses a Delphi study with discussion rounds of 231 different stakeholders (e.g., policymakers, technologists, businesspeople, academics, and civil society organizations) that are geographically spread to identify human rights issues producing AI and appropriate mitigation strategies.
Sullivan & Wamba (2022) “Moral Judgments in the Age of Artificial Intelligence” Journal of Business Ethics	<i>Responsibility attribution</i> The study outlines that moral and ethical values are complex and can vary between the several interactions between AI and humans. Hence, moral responsibility attribution depends on the relationship between perceived harm and blame judgment.	<i>Normative ethics</i> The study implicitly refers to virtue ethics as moral responsibility attribution depending on one’s inherent moral perceptions and maxims.	<i>HMA-AMA: hybrid agency</i> The authors argue that there is a shared agency between AI and humans, given their complex chain of interactions. Moral responsibility can be, therefore, attributed to companies (managers), developers/designers, or the AI system.	<i>Experiment</i> The paper builds upon a within-subject experiment with 276 Prolific workers aiming to investigate how perceived agency and perceived experiences mediate the relationship between intentional harm and blame judgment toward AI.
Sun & Ye (2023) “Moral Considerations of Artificial Intelligence” Science & Education	<i>Claim for environmental ethics</i> The authors criticize that traditional ethics focus on human beings; therefore, they use environmental ethics as it considers all entities (a more inclusive form of ethics).	<i>Meta-ethics</i> The paper refers to environmental ethics as an inclusive form of ethics (explicitly addressed).	<i>HMA: moral responsibility</i> The authors argue that morality changes with its environment, wherefore with AI, other moral issues arise compared to previous technologies. Human moral agency: “Despite many discussions about technical	<i>Conceptualization</i> The paper presents a framework for realizing moral AI in education.

			morality, the moral subject is still assigned to human designers or users” (p. 9). AI does not meet the criteria for moral agency. As moral agents are subordinate to moral patients, the AI may need to take responsibility, and the humans should be responsible for AI.	
Susser (2019) “ <i>Invisible Influence</i> ” AIES Proceedings	<i>AI perceptions</i> The study illustrates how AI influences user’s decision-making. Thus, there has to be a transparent choice architecture; otherwise, potential harms caused by AI cannot be foreseen.	<i>Meta-ethics</i> The study explicitly addresses the ethics of invisibility.	<i>HMA: limited agency</i> The author illustrates that human’s moral agency is limited when AI’s architecture is invisible. Thus, developers must engage in proactive ethical deliberation to address the potential negative consequences of invisible AI technology.	<i>Philosophical essay</i> This study is a philosophical work framing the choice architecture of AI to avoid manipulation and to increase users’ trust.
Swanepoel (2021) “ <i>The Possibility of Deliberate Norm-Adherence in AI</i> ” Ethics and Information Technology	<i>Norm-Adherence</i> This study examines the ethical implications of human-AI interaction by exploring the concepts of norm-compliance and norm-endorsement in AI systems. The research specifically focuses on the human ability to either adhere to or deliberately violate norms based on rational deliberation. For example, while an AI system might follow programmed norms in response to stimuli, such as in the case of autonomous weapons, the AI lacks the capacity to understand or deliberately violate norms. This incapacity is highlighted by Swanepoel’s assertion that AI cannot meet the “minimum requirement for moral agency” because it cannot compute the meaning of a norm and thus cannot intentionally choose to violate it (p. 160).	<i>Normative ethics</i> This study delves into the distinction between virtue ethics—where an agent acts virtuously by either following or violating norms for various reasons—and deontological ethics, which is closely aligned with Kantian ethics and focuses on norm compliance (p. 159).	<i>AMA: implicit and explicit ethical agents</i> Swanepoel builds on Moor’s (2011) differentiation between implicit and explicit ethical agents, positioning AI as implicit ethical agents programmed to promote ethical behavior, in contrast to explicit ethical agents capable of autonomous ethical judgments. The study argues that deliberate norm-adherence, which AI cannot yet achieve, is a fundamental requirement for moral agency. This contrasts sharply with human agents, who are considered responsible due to their capacity for free will and ethical reflection, grounding their personhood in the ability to choose to follow duties or not based on ethical considerations.	<i>Philosophical essay</i> This study is a philosophical exploration of the moral agency status of AI systems, concluding that current AI lacks the capacity for true moral agency.
Taeihagh (2021) “ <i>Governance of Artificial Intelligence</i> ” Policy and Society	<i>Risk management</i> The paper addresses the governance of AI, particularly in managing risks associated with unexpected and unintended consequences of AI technologies.	<i>Normative ethics</i> The study implicitly engages with deontological ethics by emphasizing the importance of governance mechanisms that seek to mitigate potential risks associated with AI automation.	<i>HMA: distributed agency</i> The author advocates for a distributed agency approach, arguing that regulatory frameworks should extend to organizational and societal institutions, such as small- and medium-sized enterprises (SMEs) and government bodies. The paper highlights the significant challenge of information asymmetry between regulators and SMEs regarding AI capabilities, underscoring the need for effective risk management strategies.	<i>Position paper</i> This study emphasizes the critical role of risk management in AI governance, advocating for a more informed and collaborative approach among stakeholders to address the ethical complexities posed by AI technologies.
Tajalli (2021) “ <i>AI Ethics and the Banality of Evil</i> ” Ethics and Information Technology	<i>The banality of evil</i> The study frames the ethics of human-AI interaction as a philosophical challenge deeply rooted in the diversity of ethical theories and principles rather than purely technological considerations	<i>Normative ethics</i> The work covers various normative ethical theories, including deontology, consequentialism, contractarianism, and virtue ethics.	<i>AMA: implicit ethical agents</i> The author argues that the ability to distinguish between good and bad—an ability requiring more than mere intelligence or programmed ethical guidelines—is essential for making ethical decisions. Consequently, AI systems, which are limited to representing specific ethical	<i>Philosophical essay</i> The study draws on Hannah Arendt’s notion of the “banality of evil” to argue that AI systems adhering strictly to programmed principles are likely to commit unethical

			principles from a single theory, cannot make genuinely ethical decisions.	acts, as they lack the capacity for true moral judgment (p. 447).
Tasiolas (2018) “First Steps Towards an Ethics of Robots and Artificial Intelligence” Journal of Practical Ethics	<i>AI perceptions on ethics of robots</i> The study provides a comprehensive overview of the ethics of robots and AI (RAI), exploring the intersection of legal regulation, social morality, and individual moral values.	<i>Normative ethics</i> The study implicitly engages with deontological ethics through its focus on regulatory standards and with virtue ethics through its emphasis on personal moral standards.	<i>HMA: moral intentionality</i> The author argues for the necessity of realistic perceptions of RAI's ethical capabilities, noting that exaggerated utopian or dystopian narratives can skew human moral values and hinder the development of appropriate ethical frameworks.	<i>Syntheses</i> The study addresses the multifaceted ethical questions surrounding RAI and calls for a balanced approach to understanding their moral agency.
Telkamp & Anderson (2022) “The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence” Journal of Practical Ethics	<i>Ethical AI based on diverse moral foundations</i> explores the concept of ethical AI through the lens of diverse moral foundations. The study emphasizes that AI systems, which adhere to norms, cannot make objectively ethical decisions due to the inherent variability in human moral judgments—a perspective aligned with moral anti-realism. The authors argue that humans must evaluate AI decisions based on three moral foundations: the purpose of AI usage, the data involved, and the consequences of AI actions.	<i>Normative ethics</i> This study addresses normative ethics by discussing the conflicts between consequentialism, deontology, and virtue ethics, highlighting the importance of shared moral understanding among humans.	<i>HMA: moral understanding</i> The authors propose that moral foundation theory can be applied to AI systems to better comprehend moral values, thereby improving the ethical alignment of AI technologies. In doing so, the authors highlight the relevance of a shared moral understanding of humans given that there can be conflicting expectations of moral values (merging human agencies is required) and given that moral foundation theory can be applied to non-human technologies that have some agency as it, at least, would help to increase the understanding of moral values by the attempt to extract and differentiate between ethical and unethical mechanisms that can be transferred to an AI system.	<i>Conceptualization</i> The study draws on the moral foundation theory of Graham (2013, 2018) to illustrate why current AI frameworks do not adequately address how different organizational stakeholders make different moral judgments concerning AI.
Terzis (2020) “Onward for the freedom of others” FAT Proceedings	<i>AI governance</i> The study critiques the current approach to AI ethics, arguing that it fails to address the fundamental issue of individual freedom. Drawing on Beauvoir's (1947, 1965) existentialist ethics, Terzis contends that ethics is an ongoing process aimed at transcending toward freedom, which cannot be captured by aggregate ethical principles or norms.	<i>Normative ethics</i> The study implicitly critiques deontological and utilitarian approaches for their tendency to impose collective norms that may restrict individual moral autonomy	<i>AMA: explicit ethical agents</i> The author argues that the freedom of will characterizes agency, wherefore morality is limited to commonly agreed values, principles, or norms that limit an individual's choices of what is morally good or bad. Talking about the notion of AI as an autonomous ethical agent having its own agency is a “fictional personification” and makes us believe that subjective values can be used to create objectives of an AI agent, as, in reality, the „ethics of AI” is the “ethics of AI” of an CEO or of a developer (p. 221).	<i>Conceptualization</i> The study draws on the work of Beauvoir (1947, 1965) by criticizing that AI ethics roundtable of different stakeholders (policymakers, lobbyists, researchers, etc.) are illusory as they only agree on a level of abstraction. Thus, the conceptual work challenges the notion of AI moral agency and emphasizes the need for ethics that respect individual freedom.
Tidjon & Khomh (2023) “The Different Faces of AI Ethics Across the World: A Principle-To-Practice Gap Analysis” IEEE Transactions on	<i>AI ethics: the principle-to-practices gap</i> The study examines the gap between AI ethics principles and their practical implementation across different countries, highlighting the variability in AI awareness and governance. Existing AI principles are not well aligned with AI practices, wherefore	<i>Normative ethics</i> The work implicitly engages with deontological ethics through its analysis of AI governance principles and contractarianism through its discussion of policy frameworks. <i>Applied ethics:</i>	<i>HMA: distributed agency</i> The authors argue that effective AI governance requires practical guidance that aligns ethical principles with real-world practices, especially in the context of the differing levels of AI awareness across nations.	<i>Syntheses</i> The study is a report on the key points for an ideal AI governance by investigating the most cited eleven global principles among studied countries (privacy, transparency,

Artificial Intelligence	practical guidance of AI principles is required.	The study further talks about how to apply ethics to AI strategizing, referring to AI-business alignment.		fairness, security, safety, responsibility, accountability, explainability, well-being, human rights, and inclusiveness.)
Tóth et al. (2022) “ <i>The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability</i> ” Journal of Business Ethics	<i>AI robot accountability</i> The study focuses on the accountability of AI robots, aligning AI ethics with business ethics as it can help guide individuals and organizations to manage ethical behavior. The study introduces a framework that differentiates between weak, moderate, and strong AI ethical agents with varying capacities for ethical decision-making.	<i>Normative & descriptive ethics</i> The authors implicitly address consequentialism by emphasizing the moral intensity of decisions, considering the magnitude and distribution of consequences.	<i>AMA: weak/strong ethical agents</i> The authors differentiated between weak AI ethical agents (not able to recognize unethical decisions), moderate AI ethical agents (imitate unethical decisions and repeat them as ethical ones), and strong ethical agents (overlook harm by pursuing other needs as well). Hence, AI can edit unethical decisions to be right, whereas other human stakeholders have to overlook the accountability of their decisions (acting as control agency). Otherwise, there is a locus of moral responsibility (from human to AI agency), especially when we talk about lethal autonomous weapon systems due to the little human involvement and the spread accountability of several actors and institutions being involved, referring to a situational moral intensity).	<i>Philosophical essay</i> The study theorizes on the ethical issues surrounding AI robots, categorizing their actions into illegal, immoral, permissible, and supererogatory, particularly concerning the ethical use of lethal autonomous weapons.
Townsend et al. (2022) “ <i>From Pluralistic Normative Principles to Autonomous-Agent Rules</i> ” Minds and Machines	<i>Institutionalized AI ethics</i> This work explores the principles of institutionalized AI ethics and the need to bridge the gap between normative ethical principles and their practical application. The authors discuss the importance of sector-specific regulations that incorporate social, legal, empathetic, and cultural considerations framed within virtues such as kindness and tolerance.	<i>Applied & normative ethics</i> The paper implicitly addresses deontological ethics through its focus on bringing ethical principles into practice, and virtue ethics through its emphasis on the role of human developers and oversight in ensuring the ethical design and deployment of AI systems.	<i>HMA: distributed agency</i> The authors mainly focus on distributed human agency (of developers) and oversight (of technical robustness, data governance, and so on).	<i>Philosophical essay</i> The study uses a case study of robot-assisted care to illustrate the challenges and opportunities in institutionalizing AI ethics. Specifically, aiming to bring the European Commission’s ethics guidelines for trustworthy AI into practice by focusing on the design of AI agents.
Unver (2023) “ <i>Rebuilding ‘Ethics’ to Govern AI: How to re-set the boundaries for the legal sector?</i> ” ICAII Proceedings	<i>Limitations of AI governance</i> The study focuses on how to apply AI ethics as accepted standards to guide moral conduct throughout the AI life cycle and legal ethics with domain-specific issues. This requires interoperability between the different stakeholders to align principles following three policy options: i) lawyers’ revision of rules/codes of conduct, ii) individual/company collaboration, and “higher-level collaboration” (e.g., between industries, universities, tech partners).	<i>Applied ethics</i> The study focuses on boundaries through policy options for the legal sector, such as applying regulation.	<i>HMA: shared responsibilities</i> The authors argue that to align the principles of different stakeholders with policies, it is required to revise existing law, to collaborate among organizational departments and across various teams, but also between societal stakeholders, such as universities and tech companies.	<i>Position paper</i> The study outlines legal ethics in the context of AI based on the three policy options.
Vakkuri et al. (2019) “ <i>Implementing Ethics in AI: Initial Results</i> ”	<i>AI ethics: ethics by design, in design, and for design</i> The study focuses on the implementation of AI ethics through design, proposing a	<i>Applied ethics</i> The study focuses on applying ethical principles in the design of AI systems.	<i>HMA: shared responsibilities</i> The authors argue that there is a chain of responsibility of several stakeholders that make decisions upon the	<i>Case Study</i> This work is an industrial multiple case study in the healthcare sector

<i>of an Industrial Multiple Case Study”</i> Computer Science	framework that includes ethics by design, in design, and for design. The authors highlight the lack of formal methods for incorporating ethical principles into software development, particularly in AI-based systems. The study argues that responsibility and accountability in AI design require a distributed agency approach, where multiple stakeholders share responsibility for ethical outcomes.		design of an AI system (internally motivated; shared responsibility; distributed agency), while to establish accountability (externally motivated), there has to be drawn connection between all those individual’s decisions and the system’s actions (hybrid accountability).	about a baseline for ethics in AI-based software development.
Vakkuri et al. (2021) “ <i>ECCOLA—A Method for Implementing Ethically Aligned AI Systems</i> ” The Journal of Systems & Software	<i>Method for ethically aligned AI</i> The study introduces ECCOLA, a method designed to help practitioners implement AI ethics in practice. The method is based on principles of practical ethics, including deontological and virtue ethics, and aims to align ethical principles with real-world practices. The study emphasizes the need for practical tools to help developers and organizations embed ethics into AI systems throughout their lifecycle. The ECCOLA method provides a structured approach to ethical AI design, incorporating stakeholder perspectives and ethical considerations into the development process. The study highlights the importance of ethical AI design and the challenges of translating ethical principles into practice.	<i>Applied ethics</i> The study introduces and tests an ethical method, called ECCOLA as a cycling action research approach to apply AI ethics to the design of AI systems.	<i>HMA: shared responsibilities</i> Following the authors, responsibility, agency, and oversights are context parts of ECCOLA as they are based on AI ethics principles.	<i>System design</i> This study is a cyclical action research approach with six stages to utilize and evaluate ECCOLA as a design method for implementing ethically aligned AI systems in different contexts.
van de Poel (2020) “ <i>Embedding Values in Artificial Intelligence (AI) Systems</i> ” Minds and Machines	<i>HCAL: Embedding values into AI</i> This study focuses on embedding values that are formulated into ethical principles as a result of design activities to respect for autonomy, non-maleficence, fairness, transparency, explainability, and accountability of AI systems.	<i>Applied ethics</i> This study uses formulated principles for the design and deployment of AI (especially driven by the EU High-Level Expert Group on AI 2019 and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019).	<i>HMA: shared responsibilities</i> The authors argue that monitoring and redesigning AI systems is important to ensure that values are kept (control agency of humans).	<i>Conceptualization</i> This work contextualizes values and what it means to embed values in AI technology.
van de Voort et al. (2015) “ <i>Refining the Ethics of Computer-made Decisions: A Classification of Moral Mediation by Ubiquitous Machines</i> ” Ethics Information Technology	<i>Moral machines</i> The study focuses on the ethics of moral decision-making within machines, which requires moral aspects in the computer’s behavior based on information processing that contains i) information acquisition, ii) information analysis, iii) decision and action selection, and iv) action implementations.	<i>Meta-ethics</i> This study discusses morality from a more general view.	<i>AMA: moral responsibility</i> The authors argue that the moral behavior of a machine can be an act (freedom to act), several acts (considering causal contributions), or the lack of an act (considering consequences). The authors further follow Noorman’s (2012) three conditions for moral responsibility: i) the agent has control over the outcomes, ii) the agent has knowledge about the outcome, iii) the agent has the choice to act. Hence, a moral decision requires human-like consciousness and free will, wherefore “machines do not only have morality in the sense of	<i>Syntheses</i> This work contains a classification of ubiquitous machine-assisted moral decisions and a classification of moral decisions of machines.

			mediating the actions of humans but that, by making their own decisions within their relations with humans, mediate morality itself.” (p. 41).	
Vanderelst & Winfield (2018) “ <i>The Dark Side of Ethical Robots</i> ” AIES Proceedings	<i>AI design</i> This study considers the pitfalls of ethical robots. Given that robots can be easily transformed from an “altruistic to an egoistic machine,” cognitive abilities are required for ethical robots to make them able to evaluate the consequences of actions and to justify their choices, which “might lead to scenarios in which we are faced with robots competing with us for the benefit of those who programmed them.” (p. 318).	<i>Normative ethics</i> This work implicitly addresses consequentialism by noting, “In fact, the difference between an ethical (i.e., seeking the most desirable outcomes for the human) robot and an aggressive (i.e., seeking the least desirable outcomes for the human) robot is a simple negation of the desirability value” (p. 321).	<i>HMA: hybrid agency</i> The authors argue that even though there are strong ethical governance policies, a robot’s ethical design can be transformed by humans as “both ethical and unethical behavior require the same cognitive machinery” (p. 320). Hence, there is a <i>hybrid moral agency</i> .	<i>Experiment</i> This study builds upon an experiment with hypothetical scenarios comparing an ethical, competitive, and aggressive robot.
Vanhée & Borit (2022) “ <i>Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence</i> ” Journal of Artificial Intelligence Research	<i>Ethical designers of AI</i> This study focuses on how to make AI designers ethical by training AI designers for ethical behavior as they develop AI systems, and given that “ethical behavior is a transferable skill, then other areas of the company activity might benefit from increased ethical practice due to the habits developed by the employees.” (p. 626)	<i>Normative ethics</i> The work addresses three ethical theories explicitly: consequentialisms, virtue ethics, and deontology, by stating, “virtue ethics, understood as an approach to normative ethics that emphasizes moral character in contrast to approaches that emphasize duties and rules (deontology) or consequences of actions (consequentialism)” (p. 619).	<i>HMA: full ethical agents</i> The authors argue that moral agency of humans that can be influenced through training moral behavior in terms of i) training moral sensitivity abilities to recognize morally sensitive aspects ii) training moral judgment abilities to choose the most sensitive course of conduct iii) training moral motivation to obtain moral duties, and iv) training moral action abilities to deploy, argue, and defend moral decisions.	<i>Conceptualization</i> The study introduces a framework for designing ethical AI using the Four Component Model of Moral Behavior (Narvaez & Rest, 1995) with four dimensions to strengthen moral behavior of individuals: i) moral sensitivity, ii) moral judgment, iii) moral motivation, and iv) moral action/implementation.
Vicã et al. (2021) “ <i>The Emperor is Naked: Moral Diplomacies and the Ethics of AI</i> ” Információs Társadalom	<i>AI ethics as a buzzword</i> This work critiques the term “AI ethics” as a buzzword that encompasses moral, legal, societal, and political considerations in AI design and deployment. The authors argue that AI ethics serves as a proxy for advancing various interests, a phenomenon often referred to as “ethics washing.” They contend that this leads to a struggle over which specific forms of power and authority AI technologies should embody (p. 92). Consequently, the authors propose that moral diplomacy is a more fitting framework to explain the ethics of AI.	<i>Descriptive ethics</i> The paper describes AI ethics as a form of diplomatic procedure, emphasizing the negotiations and compromises involved in ethical decision-making in AI. <i>Normative ethics</i> The work implicitly addresses contractarianism by discussing the bureaucratization of ethics, where ethical principles are shaped and constrained by institutional procedures and agreements.	<i>HMA: distributed agency</i> The authors argue that moral diplomacy, which seeks consensus on moral principles for AI development, leads to distributed agency, as seen in the OECD AI principles, EU principles, and UNESCO principles.	<i>Philosophical essay</i> This work is a philosophical exploration of AI ethics through the lens of moral diplomacy, analyzing current AI policies to uncover the underlying ethical dynamics.
Vidu et al. (2020) “ <i>Artificial Intelligence and the Ethical Use of Knowledge</i> ” Strategica	<i>Aligning AI strategy with business ethics</i> The authors argue that ethical guidelines for AI can help address ethical concerns and dilemmas in AI deployment. Managers are encouraged to adopt AI while aligning their strategies with principles of business ethics.	<i>Applied ethics</i> This work explores frameworks for moral decision-making in the context of aligning AI with business strategies, focusing on practical applications of ethical theory.	<i>HMA: distributed agency</i> The authors contend that responsible AI actions require a broad definition of responsibility that includes not only developers but also users, who must provide feedback to refine AI systems.	<i>Syntheses</i> The paper conducts a mapping study, identifying three main approaches through which knowledge management can address ethical challenges associated with AI.
Volkman & Gabriel (2023) “ <i>AI Moral</i> ”	<i>Augment or supplant human moral reasoning</i>	<i>Normative ethics</i> The paper explicitly engages with	<i>AMA: no moral agency</i> The authors criticize the assumption that AI can	<i>Philosophical essay</i> The work proposes a modular system of

<i>Enhancement: Upgrading the Socio-Technical System of Moral Engagement</i> Science and Engineering Ethics	This work explores how to “use AI to augment (auxiliary enhancement) or to supplant (exhaustive enhancement) human moral reasoning or judgment” (p. 11:2), questioning whether such technologies can genuinely enhance moral judgment or if they undermine human wisdom by oversimplifying complex moral decisions.	consequentialism, deontology, and virtue ethics, critiquing ideal observer theory and other alternative metaethical theories for their marginal role compared to these more established ethical frameworks.	replace human moral judgment, arguing that AI lacks the moral agency inherent in humans. They advocate for AI education that incorporates the diversity of ethical wisdom to enhance moral engagement across socio-technical systems.	multiple AI Socratic interlocutors inspired by Socratic dialogue as a means of training diverse moral traditions and pursuing moral enhancement.
Walke et al. (2023) “ <i>AI in Government: A Study on Explainability of High-Risk AI-Systems in Law Enforcement & Police Service</i> ” WI Proceedings	<i>Explainable AI</i> The paper argues that the establishment of Explainable AI (XAI) depends on the specific requirements of high-risk industries, such as law enforcement, where transparency and accountability are critical.	<i>Applied ethics</i> This work examines the AI Act of the European Commission and its requirements for XAI, applying ethical principles to ensure that AI systems are both effective and accountable. <i>Normative ethics</i> : The paper implicitly addresses deontological ethics by emphasizing the importance of acts and laws in guiding ethical AI practices.	<i>AMA-HMA: hybrid agency</i> The authors argue that while XAI is necessary for digital responsibility, additional institutional metrics are required to justify AI decisions, as complete accuracy is unattainable.	<i>Syntheses</i> This work synthesizes findings from eleven expert interviews, identifying XAI requirements for legal compliance in law enforcement and police services.
Wang et al. (2020) “ <i>Toward an Understanding of Responsible Artificial Intelligence Practices</i> ” HICSS Proceedings	<i>AI ethics and corporate social responsibility</i> The paper situates AI ethics within the framework of Corporate Social Responsibility (CSR), emphasizing that ethical AI practices are rooted in corporate culture and depend on the ethical mindset of the workforce.	<i>Applied ethics</i> The authors analyze responsible AI practices, exploring how organizations can align AI deployment with ethical principles. <i>Normative ethics</i> The work implicitly references virtue ethics, linking the ethical mindset of employees to the broader ethical culture of an organization.	<i>AMA-HMA: distributed agency</i> The authors argue that ethical behavior can be cultivated through mechanisms such as incentives, ethical AI debate clubs, and the establishment of a Chief Responsible AI Officer, all of which distribute human moral agency within the organization. While AI lacks moral agency, the practices surrounding its use can embody responsibility to some extent.	<i>Syntheses</i> The paper categorizes ten cases of responsible AI practices across various service industries in North America, Europe, and the UK.
Weber (2020) “ <i>Socio-Ethical Values and Legal Rules on Automated Platforms: The Quest for a Symbiotic Relationship</i> ” Computer Law & Security Review	<i>Linking values to regulation for trustworthy AI systems</i> This work argues that moral values encompass both socio-ethical dimensions (addressing individual and cultural needs) and legal-economic dimensions (guided by legal norms and collective limits), highlighting the intersection of ethics and law that is required to develop trustworthy AI.	<i>Applied ethics</i> The paper focuses on designing regulatory tools that align ethical principles with legal requirements, ensuring that AI systems are both lawful and ethically sound. <i>Normative ethics</i> The paper implicitly addresses virtue ethics, referring to societal values, and cultural norms, and deontology, referring to law and regulation.	<i>HMA: full ethical agents</i> The authors argue that human agency remains central in the socio-ethical dimension, emphasizing that machines cannot fully replace human moral oversight.	<i>System design</i> The work advocates for a trust-based design approach that prioritizes human values in the deployment of AI on automated platforms, contributing to the development of trustworthy AI systems.
Wehrens et al. (2023) “ <i>Ethics as Discursive Work: The Role of Ethical Framing in the Promissory Future of Data-driven Healthcare Technologies</i> ” Science,	<i>Balancing regulation and evaluation within AI practices</i> The authors describe ethics as a “balancing act” and “collective thought process,” where good AI practices are shaped not only by ethical principles or regulation but also by routines, ethical frames, and evaluations (p. 614).	<i>Applied ethics</i> The work emphasizes that ethics extends beyond abstract principles, addressing the practical realities of ethical decision-making in AI development.	<i>HMA: shared responsibilities</i> The authors argue for ethical framing as a concept that characterizes the distributed ethical responsibilities, tasks, and roles of various stakeholders in shaping AI technologies.	<i>Syntheses</i> This work outlines and analyzes three ethical framings and their implications for stakeholders’ roles and responsibilities in AI development.

Technology & Human Values				
Weiss et al. (2021) “ <i>Using the Design of Adversarial Chatbots as a Means to Expose Computer Science Students to the Importance of Ethics and Responsible Design of AI Technologies</i> ” IFIP Proceedings	<i>Responsible AI design</i> The authors argue for a novel approach to AI ethics by designing the “most unethical chatbot imaginable” “to derive principles for responsible AI design, thereby emphasizing the importance of ethical considerations in AI development. (p. 331).	<i>Applied ethics</i> This work discusses the need for responsible design and education in AI, emphasizing the ethical impact of AI technologies on society and the role of developers in ensuring ethical outcomes.	<i>HMA: hybrid moral agency</i> The paper advocates for improving curricula in computer science education to enhance the ethical design of AI, focusing on the moral agency of developers and designers.	<i>System design</i> This work describes a student project aimed at developing an adversarial chatbot as a pedagogical tool for teaching responsible AI design in computer science education.
Westerstrand (2023) “ <i>Ethics in the Intersection of AI and Democracy: The AIDEM Framework</i> ” ECIS Proceedings	<i>AI ethics and democracy</i> This work situates AI ethics within the broader field of applied ethics, addressing ethical issues that arise in the design and use of AI technologies through the lens of their relationship with democracy.	<i>Applied ethics</i> The paper applies theories of moral philosophy to analyze AI use-cases, offering a framework for resolving moral dilemmas related to AI and democracy. <i>Meta-ethics</i> The authors engage with various democratic and philosophical theories, including Rawls’ theory of justice, Nozick’s “Anarchy, State, and Utopia,” and Mouffe’s radical democracy, to explore the ethical implications of AI in democratic contexts.	<i>HMA: hybrid moral agency</i> The work argues that political deliberation and democratic processes empower citizens with agency, underscoring the responsibility of Information Systems (IS) researchers in the ethical development of AI.	<i>Position paper</i> This work develops an AIDEM (AI and democracy) framework, using philosophical theories to explore the intersection of AI ethics and democratic governance.
Willems et al. (2022) “ <i>Ethics of robotized public services: The role of robot design and its actions</i> ” Government Information Quarterly	<i>Robotic ethics</i> This study contends that the ethical considerations in robotic public services extend beyond intended outcomes, emphasizing the unintended consequences these technologies may have on human users. The authors argue that the ethical implications of robots are not just about their design but about the broader effects of their interactions with humans.	<i>Normative ethics</i> The work explicitly engages with consequentialist ethics, asserting that a robot is ethical if it can balance the rights of various users across different contexts to achieve outcomes that are ethically justified.	<i>AMA: AI perceptions</i> The study argues that the human-like appearance of robots significantly influences how people perceive their ethical behavior, thereby impacting human moral agency. The design of robots, therefore, plays a crucial role in shaping the ethical judgments of their users.	<i>Experiment</i> This work conducts a laboratory experiment with eye-tracking (n=156) and an online vignette experiment (n=1,399) to investigate the ethical evaluation of human-like robots.
Williams et al. (2021) “ <i>Teacher Perspectives on How To Train Your Robot A Middle School AI and Ethics Curriculum</i> ” AAAI Proceedings	<i>AI curricula</i> This paper discusses the often underrepresented role of ethics in computer science education, advocating for the restructuring of AI curricula to include ethical considerations. The authors argue that current AI curricula are inadequate in preparing students to handle the ethical dilemmas posed by AI technologies.	<i>Descriptive ethics</i> The study explores the ethical dilemmas that arise when using robots in educational settings, focusing on the importance of integrating ethics into the training of both teachers and students.	<i>HMA: hybrid moral agency</i> The paper emphasizes the need to equip teachers with ethical skills and awareness to better prepare students for AI technologies. It highlights the shared responsibility of teachers in delivering ethically sound AI education.	<i>Case Study</i> This work uses two case studies on K12 teachers’ perspectives to prepare middle school students for the use of robots (AI and ethics curriculum).
Wong (2021) “ <i>Ethics and Regulation of Artificial Intelligence</i> ” IFIP Proceedings	<i>Limitations of AI governance</i> This work examines the limitations of existing regulatory frameworks for AI, arguing that they are inadequate because they were developed before the advent of contemporary AI technologies. The study underscores the need for updated regulations that can	<i>Applied ethics</i> The paper provides an overview of different AI regulatory frameworks, discussing how they attempt to address ethical concerns related to AI. It calls for a more adaptive regulatory approach that can evolve	<i>HMA: shared moral responsibility</i> The research suggests that moral responsibility in AI is not confined to individual agents but is distributed across various human agents involved in the development, deployment, and regulation of AI systems.	<i>Syntheses</i> This work presents the current debates on regulatory systems surrounding AI based on an international conference.

	effectively address the ethical challenges posed by AI.	alongside technological advancement		
Wright (2020) “ <i>AI in the Law: Towards Assessing Ethical Risks</i> ” Big Data Proceedings	<i>High-level AI governance</i> This paper critiques the high-level principles that dominate ethical AI guidelines and legal frameworks, arguing that more specific, context-sensitive guidelines are necessary to effectively identify and mitigate ethical risks in AI applications, particularly in the legal domain.	<i>Normative ethics</i> The work addresses multiple ethical theories, including consequentialism, deontology, and virtue ethics. It highlights the parallels between legal frameworks, which often adopt a deontological approach, and the broader ethical landscape, which includes other valid approaches.	<i>AMA: implicit/explicit ethical agents</i> The study draws on Winfield et al.'s (2019) distinction between ethical guidelines for developers and those for the AI systems themselves. It discusses the ethical agency capabilities of machine learning systems, noting that mere compliance with user ethical constraints is insufficient.	<i>Syntheses</i> This work assesses metrics for ethics (especially testing and evaluation and a test-driven development methodology) and ethical risks for AI lawyers (competence, confidentiality, diligence, conflicts, unauthorized practice of law).
Wu et al. (2020) “ <i>Ethical Principles and Governance Technology Development of AI in China</i> ” Engineering	<i>HCAI for sustainable development of AI</i> This paper argues that ethics and governance are fundamental to the sustainable development of AI technologies. It highlights the importance of integrating ethical considerations into AI development to ensure that the technology aligns with human values and societal goals.	<i>Applied ethics</i> The study describes the ethical principles guiding AI governance in China, focusing on the country's efforts to balance technological innovation with ethical responsibility.	<i>HMA: shared moral responsibility</i> The research emphasizes the shared responsibility between government agencies and the private sector in developing and enforcing ethical guidelines for AI. It suggests that both sectors have a role in promoting the ethical use of AI technologies.	<i>Syntheses</i> This work describes the efforts of AI governance in China and the preliminary outcomes.
Yu et al. (2018) “ <i>Building Ethics into Artificial Intelligence</i> ” JCAI Proceedings	<i>Ethical decision-making for AI governance</i> This work advocates for the incorporation of ethical decision-making frameworks into AI governance to address the ethical dilemmas inherent in AI systems. It proposes a multi-tiered approach that includes individual and collective ethical decision-making frameworks as well as considerations for human-AI interactions	<i>Normative ethics</i> The study engages with consequentialism, deontology, and virtue ethics, discussing how each ethical theory can inform the development of ethical AI. It argues that AI systems should be designed to consider the moral consequences of their actions and align with fundamental ethical principles	<i>HMA: Limited human agency:</i> The research asserts that AI technologies should preserve human agency by maintaining individuals' free will. It also emphasizes the importance of distributing risks fairly among users, ensuring that AI does not disproportionately impact certain groups.	<i>Syntheses</i> This work summarizes psychological, social, and legal discussions on AI governance and proposes a taxonomy of AI governance techniques.
Zebrowski (2021) “ <i>Artificial Instinct: Lem's Robots as a Model Case for AI</i> ” Pro-Fil	<i>More nuanced models for AI ethics</i> This work criticizes the reliance on Asimov's Law in AI ethics, arguing that such fictional principles fail to capture the complexities of human morality. Instead, the author draws on Stanislaw Lem's concept of “artificial instinct” as a more nuanced model for understanding the ethical implications of AI.	<i>Meta-ethics</i> The study explores the meta-ethical implications of AI, combining metaphysical and ethical inquiries to challenge the anthropocentric biases in current AI ethics frameworks.	<i>HMA vs. AMA:</i> The work contrasts human moral agency with artificial moral agency, arguing that AI should not be judged by human standards of intelligence and ethics. Instead, it suggests that AI might better be understood through the lens of “artificial instinct,” (referring to Lem, 1986), which simulates basic human instincts rather than higher cognitive functions.	<i>Philosophical essay</i> This work describes how to use Lem's fiction about robots as a model case for AI (dealing with AI science fiction narratives, in particular, human-like AI).
Zhang & Zhang (2023) “ <i>Ethics and Governance of Trustworthy Medical Artificial Intelligence</i> ” BMC Medical Informatics and Decision Making	<i>AI governance in healthcare</i> This study explores the multi-layered governance of medical AI, arguing that effective governance must operate at ethical, legal, and regulatory levels. The authors stress the importance of stakeholder engagement in addressing the ethical and safety risks associated with medical AI.	<i>Normative ethics</i> The work draws on deontological and virtue ethics, discussing how these ethical frameworks can inform the development of trustworthy medical AI systems. It emphasizes the importance of adhering to ethical principles and norms in the governance of AI.	<i>HMA: full ethical agents</i> The study argues that while current medical AI does not possess moral status, humans remain the ultimate bearers of ethical responsibility. The governance frameworks must, therefore, ensure that human agency is preserved in the deployment of medical AI.	<i>Syntheses</i> This work uses a multidisciplinary approach that summarizes five subjects that influence the trustworthiness of medical AI (data quality, algorithmic bias, opacity, safety/security, and responsibility attribution).
Zhang et al. (2021) “ <i>Ethics and Governance of Artificial</i>	<i>Global AI governance</i> This paper builds on the definition of AI governance as the global norms and policies that ensure the beneficial design	<i>Descriptive ethics</i> The work provides a descriptive analysis of the ethical challenges in AI governance based on	<i>HMA: hybrid moral agency</i> The study underscores the expectations of AI researchers for executives and policymakers to develop	<i>Syntheses</i> This work compares the consensus and disagreements of AI researchers

<i>Intelligence: Evidence from a Survey of Machine Learning Researchers”</i> Journal of Artificial Intelligence Research	and use of AI (referring to Dafoe, 2018). The study highlights the need for a coordinated international effort to develop governance frameworks that can effectively manage the ethical challenges posed by AI.	insights from machine learning researchers. It emphasizes the importance of understanding the ethical concerns of those who are directly involved in AI development.	trustworthy AI systems and establish appropriate governance frameworks. It highlights the shared responsibility for ensuring that AI systems are ethically designed and implemented.	regarding the ethics and governance of AI.

*** Abbreviation of Conference Proceedings:**

<i>AAAI</i>	<i>AAAI Conference on Artificial Intelligence</i>
<i>ACM HCI</i>	<i>ACM on Human-Computer Interaction</i>
<i>ACIS</i>	<i>Australasian Conference on Information Systems</i>
<i>AIES</i>	<i>AAAI/ACM Conference on AI, Ethics, and Society</i>
<i>AMCIS</i>	<i>Americas Conference on Information Systems</i>
<i>CIPA</i>	<i>International Conference on Cognitive based Information Processing and Applications</i>
<i>EASE</i>	<i>International Conference on Evaluation and Assessment in Software Engineering</i>
<i>DITTET</i>	<i>International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence</i>
<i>D.GO</i>	<i>Annual International Conference on Digital Government Research</i>
<i>ECIS</i>	<i>European Conference on Information Systems</i>
<i>FAT</i>	<i>Conference on Fairness, Accountability, and Transparency</i>
<i>FACCT</i>	<i>ACM Conference on Fairness, Accountability, and Transparency</i>
<i>FICC</i>	<i>Future of Information and Communication Conference</i>
<i>HICSS</i>	<i>Hawaiian International Conference on System Sciences</i>
<i>ICAIT</i>	<i>International Conference on Applications in Information Technology</i>
<i>IHIS</i>	<i>Intelligent Human System Integration</i>
<i>ICIS</i>	<i>International Conference on Information Systems</i>
<i>IEEE</i>	<i>Institute of Electrical and Electronics Engineers</i>
<i>IFIP</i>	<i>International Federation for Information Processing</i>
<i>IJCIME</i>	<i>International Joint Conference on Information, Media and Engineering</i>
<i>IMES</i>	<i>International Conference: Innovation Management, Entrepreneurship and Sustainability</i>
<i>IRIS</i>	<i>Information Systems Research Seminar in Scandinavia</i>
<i>JCAI</i>	<i>International Joint Conference on Artificial Intelligence</i>
<i>LNAI</i>	<i>Lecture Notes in Artificial Intelligence</i>
<i>PACM</i>	<i>Proceedings of the ACM</i>
<i>SCIS</i>	<i>Scandinavian Conference on Information Systems</i>
<i>WI</i>	<i>Wirtschaftsinformatik</i>

Appendix VI. Descriptive Statistics and Meta Information on the Final Sample

Table 9 Descriptive statistics and meta information

Category	Meta Information	Statistics	Articles
Database	ACM	N = (26/151) 17,21 %	Agbese et al., 2023; Bleher & Braun, 2023; Burema et al., 2023; Chaput et al., 2021; Colmenarejo et al., 2022; Cook, 2023; Cooper et al., 2022; Corvite et al., 2023; Dyrkolbotn et al., 2018; Eicher et al., 2018; Estrada, 2018; Fabiano, 2019; Ha, 2022; Hawkins & Mittelstadt, 2023; Henderson, 2019; Hooker & Kim, 2018; Lim & Kwon, 2021; McDonald & Pan, 2020; Rismani & Moon, 2023; Siapka, 2022; Susser, 2019; Terzis, 2020; Unver, 2023; Vanderelst & Winfield, 2018; Vanhée & Borit, 2022; B. Zhang et al., 2021
	AIS	N= (18/151) 11,92 %	Bilal et al., 2021; Emdad et al., 2023; Figueras et al., 2022; Ghatar et al., 2023; Greene et al., 2019; Jantunen et al., 2021; Mäntymäki et al., 2023; Mayer et al., 2020; Minkinen & Mäntymäki, 2023; Polyviou & Zamani, 2022; Schlimbach & Khosrawi-Rad, 2022; Schneider et al., 2023; Seo & Thorson, 2022; Seppälä et al., 2021; Seymour, 2018; Walke et al., 2023; Wang et al., 2020; Westerstrand, 2023
	BSC	N= (2/151) 1,32 %	Auld et al., 2022; Gerdes, 2022
	IEEE	N= (3/151) 1,98 %	Kowch, 2019; Ong, 2021; Tidjon & Khomh, 2023
	OVID	N= (15/151) 9,93 %	Ashrafian, 2015; Boyles, 2018; Buhmann & Fieseler, 2023; Dodig Crnkovic & Çürüklü, 2012; Erman & Furendal, 2022; Etzioni & Etzioni, 2017; Kornilaev, 2021; Lloyd, 1985; Manna & Nath, 2021; Munoko et al., 2020; Sullivan & Fosso Wamba, 2022; Tasioulas, 2018; Telkamp & Anderson, 2022; Van De Voort et al., 2015; Zebrowski, 2021
	SciDir	N= (1/151) 0,66 %	Stahl et al., 2023
	WoS	N= (85/151) 56,29 %	Allen et al., 2000; Arnold & Scheutz, 2018; Ashrafian, 2015; Badea & Artus, 2022; Bankins & Formosa, 2023; Behdadi & Munthe, 2020; Bogosian, 2017; Bryson, 2018; Buruk et al., 2020; Carneiro & Veloso, 2022; Champagne & Tonkens, 2023; Coeckelbergh, 2020; Conitzer et al., 2017; Constantinescu et al., 2021; Córdova & Vicari, 2022; Farisco et al., 2020; Forsyth et al., 2021; Fukuda-Parr & Gibbons, 2021; Gahnberg, 2021; Génova et al., 2023; Giroux et al., 2022; Gómez De Ágreda, 2020; Hagendorff, 2020; Hauer, 2020, 2022; Horváth, 2022; Jobin et al., 2019; Katznelson & Gerke, 2021; Kenny et al., 2021; Khan et al., 2022; Kieslich et al., 2022; Kim et al., 2018; Knox, 2023; Köbis et al., 2021; Koniakou, 2023; Kučera, 2019; Lara, 2021; Li, 2021; Mabaso, 2021; Martin, 2019; Martinho et al., 2021; Millar, 2016; Morley et al., 2021; Neubert & Montañez, 2020; Nevanperä et al., 2021; Niederman & Baker, 2023; Obi & Gray, 2023; Palmer & Schwan, 2024; Pansoni et al., 2023; Parlangeli et al., 2020; Pruski, 2023; Raji et al., 2021; Roberts et al., 2023; Roberts & Montoya, 2022b; Rodgers et al., 2023; Saariluoma & Leikas, 2020; Serafimova, 2020; Shank et al., 2019; Shevlin, 2021; Shih et al., 2021; Shneiderman, 2020; Siau & Wang, 2020; Smith & Vickers, 2021; Stahl et al., 2023; Sun & Ye, 2023; Swanepoel, 2021; Taeihagh, 2021; Tajalli, 2021; Tóth et al., 2022; Townsend et al., 2022; Vakkuri et al., 2019, 2021; Van De Poel, 2020b; Vicá et al., 2021; Vidu et al., 2020; Volkman & Gabriels, 2023; Weber, 2020; Wehrens et al., 2023; Wright, 2020; Wu et al., 2020; Yu et al., 2018; Zhang & Zhang, 2023
Publication Type	Journal	N= (92/151) 60,93 %	Allen et al., 2000; Arnold & Scheutz, 2018; Ashrafian, 2015, 2015; Auld et al., 2022; Bankins & Formosa, 2023; Behdadi & Munthe, 2020; Bleher & Braun, 2023; Bogosian, 2017; Boyles, 2018; Bryson, 2018; Buhmann & Fieseler, 2023; Buruk et al., 2020; Coeckelbergh, 2020; Constantinescu et al., 2021; Córdova & Vicari, 2022; Dodig Crnkovic & Çürüklü, 2012; Erman & Furendal, 2022; Etzioni & Etzioni, 2017; Farisco et al., 2020; Figueras et al., 2022; Fukuda-Parr & Gibbons, 2021; Gahnberg, 2021; Génova et al., 2023; Gerdes, 2022; Giroux et al., 2022; Gómez De Ágreda, 2020; Hagendorff, 2020; Hauer, 2020, 2022; Horváth, 2022; Jantunen et al., 2021; Jobin et al., 2019; Katznelson & Gerke, 2021; Kenny et al., 2021; Khan et al., 2022; Kieslich et al., 2022; Knox, 2023; Köbis et al., 2021; Koniakou, 2023; Kornilaev, 2021; Kučera, 2019; Lara, 2021; Li, 2021; Lloyd, 1985; Mabaso, 2021; Manna & Nath, 2021; Martin, 2019; Martinho et al., 2021; Millar, 2016; Morley et al., 2021; Munoko et al., 2020; Neubert & Montañez, 2020; Niederman & Baker, 2023; Palmer & Schwan, 2024; Pruski, 2023; Roberts et al., 2023; Roberts & Montoya, 2022; Rodgers et al., 2023; Serafimova, 2020; Shank et al., 2019; Shevlin, 2021; Shih et al., 2021; Shneiderman, 2020; Siau & Wang, 2020; Smith

			& Vickers, 2021; Stahl et al., 2022, 2023; Sullivan & Fosso Wamba, 2022; Sun & Ye, 2023; Swanepoel, 2021; Taeihagh, 2021; Tajalli, 2021; Tasioulas, 2018; Telkamp & Anderson, 2022; Tidjon & Khomh, 2023; Tóth et al., 2022; Townsend et al., 2022; Vakkuri et al., 2019, 2021; Van De Poel, 2020; Van De Voort et al., 2015; Vanhée & Borit, 2022; Vică et al., 2021; Volkman & Gabriels, 2023; Weber, 2020; Wehrens et al., 2023; Willems et al., 2022; Wong, 2021; Wu et al., 2020; Zebrowski, 2021; B. Zhang et al., 2021; Zhang & Zhang, 2023
	Conference	N= (59/151) 39,07 %	Agbese et al., 2023; Badea & Artus, 2022; Bilal et al., 2021; Burema et al., 2023; Carneiro & Veloso, 2022; Champagne & Tonkens, 2023; Chaput et al., 2021; Colmenarejo et al., 2022; Conitzer et al., 2017; Cook, 2023; Cooper et al., 2022; Corvite et al., 2023; Dyrkolbotn et al., 2018; Eicher et al., 2018; Emdad et al., 2023; Estrada, 2018; Fabiano, 2019; Forsyth et al., 2021; Ghatar et al., 2023; Greene et al., 2019; Ha, 2022; Hawkins & Mittelstadt, 2023; Henderson, 2019; Hooker & Kim, 2018; Jantunen et al., 2021; Kim et al., 2018; Kowch, 2019; Lim & Kwon, 2021; Mäntymäki et al., 2023; Nevanperä et al., 2021; Obi & Gray, 2023; Ong, 2021; Pansoni et al., 2023; Parlangeli et al., 2020; Polyviou & Zamani, 2022; Raji et al., 2021; Rismani & Moon, 2023; SaariLuoma & Leikas, 2020; Schlimbach & Khosrawi-Rad, 2022; Schneider et al., 2023; Seo & Thorson, 2022; Seppälä et al., 2021; Seymour, 2018; Siapka, 2022; Susser, 2019; Terzis, 2020; Unver, 2023; Vanderelst & Winfield, 2018; Vidu et al., 2020; Walke et al., 2023; Wang et al., 2020; Weiss et al., 2021; Westerstrand, 2023; Williams et al., 2021; Wright, 2020; Yu et al., 2018
Quality	High (Impact Factor above 5, Ranking A*, A)	N= (30/151) 19,87 %	Bankins & Formosa, 2023; Behdadi & Munthe, 2020; Bogosian, 2017; Buhmann & Fieseler, 2023; Coeckelbergh, 2020; Etzioni & Etzioni, 2017; Giroux et al., 2022; Gómez De Ágreda, 2020; Hagendorff, 2020; Jobin et al., 2019; Knox, 2023; Koniakou, 2023; Kornilav, 2021; Martin, 2019; Martinho et al., 2021; Millar, 2016; Morley et al., 2021; Munoko et al., 2020; Neubert & Montañez, 2020; Niederman & Baker, 2023; Pruski, 2023; Stahl et al., 2023; Sullivan & Fosso Wamba, 2022; Taeihagh, 2021; Telkamp & Anderson, 2022; Tóth et al., 2022; Townsend et al., 2022; Van De Poel, 2020; Willems et al., 2022; Wu et al., 2020
	Middle (Impact Factor 1-5, Ranking B-D)	N= (83/151) 54,96 %	Allen et al., 2000; Arnold & Scheutz, 2018; Ashrafian, 2015, 2015; Auld et al., 2022; Bleher & Braun, 2023; Bryson, 2018; Burema et al., 2023; Buruk et al., 2020; Champagne & Tonkens, 2023; Chaput et al., 2021; Conitzer et al., 2017; Constantinescu et al., 2021; Cook, 2023; Corvite et al., 2023; Dodig Crnkovic & Cürüklü, 2012; Eicher et al., 2018; Emdad et al., 2023; Fabiano, 2019; Farisco et al., 2020; Figueras et al., 2022; Fukuda-Parr & Gibbons, 2021; Gahnberg, 2021; Génova et al., 2023; Ghatar et al., 2023; Greene et al., 2019; Ha, 2022; Hauer, 2020, 2022; Hawkins & Mittelstadt, 2023; Henderson, 2019; Hooker & Kim, 2018; Horváth, 2022; Katznelson & Gerke, 2021; Kenny et al., 2021; Khan et al., 2022; Kieslich et al., 2022; Köbis et al., 2021; Kučera, 2019; Lara, 2021; Li, 2021; Mabaso, 2021; Mäntymäki et al., 2023; Mayer et al., 2020; Ong, 2021; Polyviou & Zamani, 2022; Raji et al., 2021; Rismani & Moon, 2023; Roberts et al., 2023; Schlimbach & Khosrawi-Rad, 2022; Schneider et al., 2023; Seo & Thorson, 2022; Seppälä et al., 2021; Serafimova, 2020; Seymour, 2018; Shank et al., 2019; Shevlin, 2021; Shih et al., 2021; Shneiderman, 2020; Siau & Wang, 2020; Smith & Vickers, 2021; Stahl et al., 2023; Sun & Ye, 2023; Susser, 2019; Swanepoel, 2021; Tajalli, 2021; Tidjon & Khomh, 2023; Vakkuri et al., 2021; Van De Voort et al., 2015; Vanderelst & Winfield, 2018; Vanhée & Borit, 2022; Volkman & Gabriels, 2023; Walke et al., 2023; Wang et al., 2020; Weber, 2020; Wehrens et al., 2023; Weiss et al., 2021; Westerstrand, 2023; Williams et al., 2021
	Low (Impact Factor below 1, Ranking lower than D)	N= (23/151) 15,23 %	Agbese et al., 2023; Badea & Artus, 2022; Bilal et al., 2021; Colmenarejo et al., 2022; Córdova & Vicari, 2022; Dyrkolbotn et al., 2018; Erman & Furendal, 2022; Lloyd, 1985; Manna & Nath, 2021; Minkinen & Mäntymäki, 2023; Nevanperä et al., 2021; Palmer & Schwan, 2024; Parlangeli et al., 2020; Roberts & Montoya, 2022; Rodgers et al., 2023; Siapka, 2022; Tasioulas, 2018; Unver, 2023; Vakkuri et al., 2019; Vică et al., 2021; Vidu et al., 2020; Wong, 2021
	No ranking	N= (15/151) 9,93%	Boyles, 2018; Carneiro & Veloso, 2022; Cooper et al., 2022; Forsyth et al., 2021; Gerdes, 2022; Jantunen et al., 2021; Kim et al., 2018; Kowch, 2019; Lim & Kwon, 2021; McDonald & Pan, 2020; Obi & Gray, 2023; Pansoni et al., 2023; SaariLuoma & Leikas, 2020; Terzis, 2020; Zebrowski, 2021
	Top IS *according to our analysis which can be found there: https://ethicaltheories.github.io	N= (13/151) 8,60%	Emdad et al., 2023; Ghatar et al., 2023; Greene et al., 2019; Mäntymäki et al., 2023; Mayer et al., 2020; Polyviou & Zamani, 2022; Schlimbach & Khosrawi-Rad, 2022; Schneider et al., 2023; Seo & Thorson, 2022; Seppälä et al., 2021; Wang et al., 2020; Westerstrand, 2023; Zhang & Zhang, 2023

Method /Analysis Type	Bibliometric Analysis	N= (1/151) 0,66%	Munoko et al., 2020
	Descriptive	N= (2/151) 1,32%	Henderson, 2019; B. Zhang et al., 2021
	Case Study with Interviews	N= (7/151) 4,64%	Ha, 2022; Henderson, 2019; Neubert & Montañez, 2020; Townsend et al., 2022; Vakkuri et al., 2019; Wang et al., 2020; Weiss et al., 2021
	Grounded Theory	N= (2/151) 1,32%	Vakkuri et al., 2019; Walke et al., 2023
	Delphi Study	N= (1/151) 0,66 %	Stahl et al., 2023
	Mixed	N= (1/151) 0,66 %	Corvite et al., 2023
	Qualitative Analysis	N= (34/151) 22,52%	Agbese et al., 2023; Bilal et al., 2021; Burema et al., 2023; Buruk et al., 2020; Champagne & Tonkens, 2023; Cooper et al., 2022; Emdad et al., 2023; Figueras et al., 2022; Ghatar et al., 2023; Gómez De Ágreda, 2020; Ha, 2022; Hagendorff, 2020; Jobin et al., 2019; Kenny et al., 2021; Kim et al., 2018; Lim & Kwon, 2021; Mäntymäki et al., 2023; Mayer et al., 2020; Minkkinen & Mäntymäki, 2023; Neubert & Montañez, 2020; Obi & Gray, 2023; Pansoni et al., 2023; Polyviou & Zamani, 2022; Raji et al., 2021; Rismani & Moon, 2023; Roberts & Montoya, 2022; Seppälä et al., 2021; Stahl et al., 2022; Tidjon & Khomh, 2023; Townsend et al., 2022; Wang et al., 2020; Wehrens et al., 2023; Weiss et al., 2021; Williams et al., 2021
	Quantitative Analysis / Descriptive Statistics	N= (8/151) 5,30%	Chaput et al., 2021; Giroux et al., 2022; Kieslich et al., 2022; Parlangeli et al., 2020; Schneider et al., 2023; Shank et al., 2019; Sullivan & Fosso Wamba, 2022; Willems et al., 2022
	Other (mainly philosophical essay)	N= (102/151) 67,55%	Allen et al., 2000; Arnold & Scheutz, 2018; Ashrafian, 2015, 2015; Auld et al., 2022; Badea & Artus, 2022; Bankins & Formosa, 2023; Behdadi & Munthe, 2020; Bleher & Braun, 2023; Bogosian, 2017; Boyles, 2018; Bryson, 2018; Buhmann & Fieseler, 2023; Carneiro & Veloso, 2022; Coeckelbergh, 2020; Colmenarejo et al., 2022; Conitzer et al., 2017; Constantinescu et al., 2021; Cook, 2023; Córdova & Vicari, 2022; Dodig Crnkovic & Çürüklü, 2012; Dyrkolbotn et al., 2018; Eicher et al., 2018; Erman & Furendal, 2022; Estrada, 2018; Etzioni & Etzioni, 2017; Fabiano, 2019; Farisco et al., 2020; Forsyth et al., 2021; Fukuda-Parr & Gibbons, 2021; Gahnberg, 2021; Génova et al., 2023; Gerdes, 2022; Greene et al., 2019; Hauer, 2020, 2022; Hawkins & Mittelstadt, 2023; Hooker & Kim, 2018; Horváth, 2022; Jantunen et al., 2021; Katznelson & Gerke, 2021; Khan et al., 2022; Knox, 2023; Köbis et al., 2021; Koniakou, 2023; Kornilaev, 2021; Kowch, 2019; Kučera, 2019; Lara, 2021; Li, 2021; Lloyd, 1985; Mabaso, 2021; Manna & Nath, 2021; Martin, 2019; Martinho et al., 2021; McDonald & Pan, 2020; Millar, 2016; Morley et al., 2021; Nevanperä et al., 2021; Niederman & Baker, 2023; Ong, 2021; Palmer & Schwan, 2024; Pruski, 2023; Roberts et al., 2023; Rodgers et al., 2023; Saarihuoma & Leikas, 2020; Schlimbach & Khosrawi-Rad, 2022; Seo & Thorson, 2022; Serafimova, 2020; Seymour, 2018; Shevlin, 2021; Shih et al., 2021; Shneiderman, 2020; Siapka, 2022; Siau & Wang, 2020; Smith & Vickers, 2021; Sun & Ye, 2023; Susser, 2019; Swanepoel, 2021; Taeihagh, 2021; Tajalli, 2021; Tasioulas, 2018; Telkamp & Anderson, 2022; Terzis, 2020; Tóth et al., 2022; Unver, 2023; Vakkuri et al., 2021; Van De Poel, 2020; Van De Voort et al., 2015; Vanderelst & Winfield, 2018; Vanhée & Borit, 2022; Vičá et al., 2021; Vidu et al., 2020; Volkman & Gabriels, 2023; Weber, 2020; Westerstrand, 2023; Wong, 2021; Wright, 2020; Wu et al., 2020; Yu et al., 2018; Zebrowski, 2021; Zhang & Zhang, 2023
Research Context	(Higher) Education	N= (13/151) 8,61%	Champagne & Tonkens, 2023; Córdova & Vicari, 2022; Eicher et al., 2018; Forsyth et al., 2021; Henderson, 2019; Kučera, 2019; McDonald & Pan, 2020; Parlangeli et al., 2020; Schlimbach & Khosrawi-Rad, 2022; Shih et al., 2021; Sun & Ye, 2023; Weiss et al., 2021; Williams et al., 2021
	Bioethics	N= (2/151) 1,32 %	Hauer, 2020; Lara, 2021
	Firms / Consulting	N= (6/151) 3,97 %	Corvite et al., 2023; Munoko et al., 2020; Neubert & Montañez, 2020; Rismani & Moon, 2023; Shank et al., 2019; Stahl et al., 2023
	Healthcare	N= (12/151) 7,95%	Buruk et al., 2020; Emdad et al., 2023; Katznelson & Gerke, 2021; Kenny et al., 2021; Köbis et al., 2021; Palmer & Schwan, 2024; Pruski, 2023; Schneider et al., 2023; Shevlin, 2021; Vakkuri et al., 2019; Wehrens et al., 2023; Zhang & Zhang, 2023
	Military	N= (2/151) 1,32 %	Ashrafian, 2015; Gómez De Ágreda, 2020
	Public Organizations / Governance / Legal Sector	N= (14/151) 9,27%	Burema et al., 2023; Erman & Furendal, 2022; Figueras et al., 2022; Giroux et al., 2022; Koniakou, 2023; Mäntymäki et al., 2023; Roberts et al., 2023; Roberts & Montoya, 2022; Shneiderman, 2020; Unver,

			2023; Vică et al., 2021; Walke et al., 2023; Willems et al., 2022; Wright, 2020
	Software Engineering / Robotics	N= (4/151) 2,65%	Agbese et al., 2023; Allen et al., 2000; Mayer et al., 2020; Townsend et al., 2022
	No specific context	N= (98/151) 64,9 %	Arnold & Scheutz, 2018; Ashrafian, 2015; Auld et al., 2022; Badea & Artus, 2022; Banks & Formosa, 2023; Behdadi & Munthe, 2020; Bilal et al., 2021; Bleher & Braun, 2023; Bogosian, 2017; Boyles, 2018; Bryson, 2018; Buhmann & Fieseler, 2023; Carneiro & Veloso, 2022; Chaput et al., 2021; Coeckelbergh, 2020; Colmenarejo et al., 2022; Conitzer et al., 2017; Constantinescu et al., 2021; Cook, 2023; Cooper et al., 2022; Dodig Crnkovic & Çürüklü, 2012; Dyrkolbotn et al., 2018; Estrada, 2018; Etzioni & Etzioni, 2017; Fabiano, 2019; Farisco et al., 2020; Fukuda-Parr & Gibbons, 2021; Gahnberg, 2021; Génova et al., 2023; Gerdes, 2022; Ghatat et al., 2023; Greene et al., 2019; Ha, 2022; Hagendorff, 2020; Hauer, 2022; Hawkins & Mittelstadt, 2023; Hooker & Kim, 2018; Horváth, 2022; Jantunen et al., 2021; Jobin et al., 2019; Khan et al., 2022; Kieslich et al., 2022; Kim et al., 2018; Knox, 2023; Kornilae, 2021; Kowch, 2019; Li, 2021; Lim & Kwon, 2021; Lloyd, 1985; Mabaso, 2021; Manna & Nath, 2021; Martin, 2019; Martinho et al., 2021; Millar, 2016; Minkinen & Mäntymäki, 2023; Morley et al., 2021; Nevanperä et al., 2021; Niederman & Baker, 2023; Obi & Gray, 2023; Ong, 2021; Pansoni et al., 2023; Polyviou & Zamani, 2022; Raji et al., 2021; Rodgers et al., 2023; Saariluoma & Leikas, 2020; Seo & Thorson, 2022; Seppälä et al., 2021; Serafimova, 2020; Seymour, 2018; Siapka, 2022; Siau & Wang, 2020; Smith & Vickers, 2021; Stahl et al., 2022; Sullivan & Fosso Wamba, 2022; Susser, 2019; Swanepoel, 2021; Taeihagh, 2021; Tajalli, 2021; Tasioulas, 2018; Telkamp & Anderson, 2022; Terzis, 2020; Tidjon & Khomh, 2023; Tóth et al., 2022; Vakkuri et al., 2021; Van De Poel, 2020; Van De Voort et al., 2015; Vanderelst & Winfield, 2018; Vanhée & Borit, 2022; Vidu et al., 2020; Volkman & Gabriels, 2023; Wang et al., 2020; Weber, 2020; Westerstrand, 2023; Wong, 2021; Wu et al., 2020; Yu et al., 2018; Zebrowski, 2021; B. Zhang et al., 2021
Result / Contribution Type	Propositions / Avenues	N= (28/151) 18,54 %	Allen et al., 2000; Champagne & Tonkens, 2023; Coeckelbergh, 2020; Martin, 2019; Saariluoma & Leikas, 2020; Shih et al., 2021; Zebrowski, 2021
	Framework / Classification / Synthesis / Conceptualization	N= (64/151) 42,38 %	Agbese et al., 2023; Ashrafian, 2015; Banks & Formosa, 2023; Bleher & Braun, 2023; Buhmann & Fieseler, 2023; Burema et al., 2023; Buruk et al., 2020; Colmenarejo et al., 2022; Emdad et al., 2023; Farisco et al., 2020; Forsyth et al., 2021; Fukuda-Parr & Gibbons, 2021; Ghatat et al., 2023; Gómez De Ágreda, 2020; Greene et al., 2019; Horváth, 2022; Jobin et al., 2019; Katznelson & Gerke, 2021; Kenny et al., 2021; Kim et al., 2018; Koniakou, 2023; Lim & Kwon, 2021; Lloyd, 1985; Mabaso, 2021; Mäntymäki et al., 2023; McDonald & Pan, 2020; Minkinen & Mäntymäki, 2023; Munoko et al., 2020; Neubert & Montañez, 2020; Niederman & Baker, 2023; Ong, 2021; Palmer & Schwan, 2024; Pansoni et al., 2023; Polyviou & Zamani, 2022; Pruski, 2023; Raji et al., 2021; Rismani & Moon, 2023; Rodgers et al., 2023; Schneider et al., 2023; Seppälä et al., 2021; Seymour, 2018; Shneiderman, 2020; Siapka, 2022; Siau & Wang, 2020; Smith & Vickers, 2021; Stahl et al., 2023; Sullivan & Fosso Wamba, 2022; Sun & Ye, 2023; Taeihagh, 2021; Telkamp & Anderson, 2022; Tidjon & Khomh, 2023; Townsend et al., 2022; Unver, 2023; Van De Poel, 2020; Van De Voort et al., 2015; Vică et al., 2021; Walke et al., 2023; Wang et al., 2020; Weber, 2020; Wehrens et al., 2023; Westerstrand, 2023; Wong, 2021; Wright, 2020; Wu et al., 2020
	Hypotheses / Research Model Validation / Testing	N= (10/151) 6,62%	Bilal et al., 2021; Corvite et al., 2023; Giroux et al., 2022; Kieslich et al., 2022; Martinho et al., 2021; Obi & Gray, 2023; Parlangeli et al., 2020; Shank et al., 2019; Stahl et al., 2023; Willems et al., 2022
	Theory-Development	N= (19/151) 12,58%	Ashrafian, 2015; Bryson, 2018; Constantinescu et al., 2021; Cook, 2023; Erman & Furendal, 2022; Etzioni & Etzioni, 2017; Figueras et al., 2022; Ha, 2022; Hauer, 2020; Kornilae, 2021; Li, 2021; Manna & Nath, 2021; Millar, 2016; Morley et al., 2021; Seo & Thorson, 2022; Serafimova, 2020; Shevlin, 2021; Swanepoel, 2021; Terzis, 2020
	(Practical/Design) Guidelines	N= (18/151) 11,92%	Arnold & Scheutz, 2018; Boyles, 2018; Chaput et al., 2021; Eicher et al., 2018; Génova et al., 2023; Gerdes, 2022; Hagendorff, 2020; Jantunen et al., 2021; Knox, 2023; Lara, 2021; Mayer et al., 2020; Nevanperä et al., 2021; Schlimbach & Khosrawi-Rad, 2022; Susser, 2019; Tóth et al., 2022; Vakkuri et al., 2021; Vanhée & Borit, 2022; Vidu et al., 2020; Williams et al., 2021
	Description	N= (2/151) 1,32%	Kučera, 2019; Vidu et al., 2020
	Experiment	N= (1/151) 0,66 %	Vanderelst & Winfield, 2018

	Discourse	N= (7/151) 4,64 %	Allen et al., 2000; Champagne & Tonkens, 2023; Coeckelbergh, 2020; Martin, 2019; Saariluoma & Leikas, 2020; Shih et al., 2021; Zebrowski, 2021
	Computational Framework	N= (2/151) 1,32 %	Bogosian, 2017; Hooker & Kim, 2018

Appendix VII. Coding Scheme

Table 10 Coding scheme

Category	Explanation	Codes	Coding Rules	Example
Normative ethics (Ethical theory)	Aims to answer questions about what is morally right or wrong to guide people on how to act.	Deontology/ Deontological ethics (<i>deductive</i>)	- Ethics is rooted in duty and obligation - Actions can be inherently moral or immoral regardless of their outcomes - An action is morally right if it can be universalized as general law - Similar concepts: deontology, duty ethics, ethical logic, Kant's categorical imperative, bottom-up proceduralism	"Norm-compliance refers to Kant's notion of acting in accordance with the norm, and norm-endorsement (necessary for deliberate norm-adherence) is to act from that norm through embracing it as one's own." (Swanepoel, 2021, p. 159)
		Virtue ethics (<i>deductive</i>)	- Ethics is rooted in one's virtue and thoughts leading to an action - The maxima behind an action can be inherently moral or immoral (regardless of the action itself and its consequences) - Similar concepts: ethical motivation, inherent ethical values, moral virtues, Nicomachean Ethics, one's personal/virtuous character (e.g., courage, honesty)	"In the case of this study, a list of virtues could be the representation of the universe of virtue ethics concourse." (Bilal et al., 2021, p. 5)
		Consequentialism (<i>deductive</i>)	- Ethics is rooted in judging the outcomes of actions - The consequence of an action determines whether it is morally right or wrong - An action is morally right if it generates the greatest benefits for the greatest number of people - Similar concepts: utilitarianism, utilitarian ethics, Mill's greatest happiness principle	"Some hold that the fundamental moral norm is the principle of utility, which defines right actions and policies in terms of maximizing aggregate good consequences, while others hold that certain kinds of actions are justifiable even if a particular token of the type would maximize aggregate good." (Allen et al., 2000, p. 252)
		Contractarianism (<i>deductive</i>)	- Related to deontological ethics (law) - Societal, industrial, organizational policies that determine ethical duties for certain groups of people (non-compliance can lead to sanctions) - Claims that moral norms derive their normative force from the idea of contract or mutual agreement - Similar concepts: norm-endorsement, social contract theory	"These contracts act like a law for the individuals when they violate these laws they get punishment or send to prison. These contracts are made on the mutual agreement of the society members based on moral and political behavior." (Simpson, 1992, p. 3)
		Theory of Justice (<i>deductive</i>)	- Ethics is rooted in the idea that every individual has an equal right to basic liberties (right to opportunities, equal chance) - Actions of citizens and representatives are judged based on their distributive justice in society (e.g., when and why democracy is morally desirable or not) - Moral theory alternative to Utilitarianism (consequentialism) - Examples: Theory of Justice and Anarchy, State, and Utopia (Nozick, 1974), Normative Democratic Theory mainly based on Rawls (1999)	"These ethical implications are then reflected towards the ideal of deliberative democracy, which draws inspiration from Rawls's theory of justice. It is thus assumed that alignments or dissonances with Rawls's theory of justice would have implications for deliberative democracy." (Westerstrand, 2023, p. 2)
Meta-ethics (Ethical theory)	Concerns ethics and morality from a philosophical, epistemological, or ontological view and the nature of right and wrong.	Feminist meta-ethics (<i>inductive</i>)	- An approach to ethics that builds on the belief that traditionally ethical theorizing has undervalued and/or underappreciated women's moral experience - Similar concepts: feminist values, feminine ethics, feminist code of ethics	"In like manner, feminist metaethics is distinguished not only by its key topics and their framing but also by its tools and methods. It draws together research strands, which, despite their admittedly wide diversity, share an opposition to traditional metaethics." (Siapka, 2022, p. 667)

		Environmental ethics (<i>inductive</i>)	• Conceptual foundations of environmental values and concrete issues surrounding societal attitudes, actions, and policies to protect and sustain biodiversity and ecological systems. • Similar concepts: ethics of sustainability, sustainable ethics	“During the process of development, environmental ethics goes further and considers all life, while existentialism expands the scope of moral consideration by combining with anything that already, actual, or possible existence, and provides the most universal and comprehensive definition of inclusive ethics in this way.” (Sun & Ye, 2023, p. 14)
		Ethics of invisibility (<i>inductive</i>)	• Ethics that is about the moral machinery of our own hearts (that no one sees but self) and examines invisible social identities • Similar concepts: invisible ethics	“Finally, because the designers of these systems are highly incentivized to subvert user interests, and because the invisible nature of invisible influence renders it exceedingly difficult to detect, demands for radical transparency and seamless design must be accompanied by stringent accountability measures.” (Susser, 2019, p. 407)
	Descriptive ethics (Ethical theory) Context	Trolley problem (<i>inductive</i>)	• Experimental studies involving thought experiments in ethics studying stylized ethical dilemmas (weighting two moral maxims) • Similar concepts: trolley dilemma, moral dilemma study, moral decision-making	“While humans are always attributed greater fault than AIs, in situations where a human is considered to be greatly at fault, an AI is deemed to be relatively fault-free. In situations where the human is attributed less fault, an AI committing the same act is attributed more fault than it was in the situations where the human was greatly at fault. This finding suggests that the types of moral violations deemed most severely wrong for humans are the least wrong for an AI to commit.” (Shank et al., 2019, p. 659)
		The Ethical Turing Test (<i>inductive</i>)	• Test in which examples of morally actions of humans and a machine are given, wherefore judgments on whether the action is less moral or not should be made • Similar concepts: comparative Moral Turing Test of Allan et al. (2000)	“A Moral Turing Test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human ‘interrogators’ cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent.” (Allen et al., 2000, p. 354)
	Concerns the design and application of ethical principles and guidelines to deal with real-world moral dilemmas and issues.	AI curricula (<i>inductive</i>)	• Curricula, syllabus for educating humans • Related to deontological ethics (guidelines) • Related concepts: K-12, K-9 education, AI literacy	“We developed the How to Train Your Robot: AI and Ethics Curriculum for middle school teachers who want to introduce AI to their students.” (Williams et al., 2021, p. 15678)
		Business ethics (<i>deductive</i>)	• Guidance of organizational stakeholders to foster ethical behavior • Related to deontological ethics (guidelines) • Related concepts: Corporate Social Responsibility (CSR), Corporate Digital Responsibility (CDR)	“However, normative business ethics studies moral principles and develops guidance for resolving individual/institutional moral dilemmas and market design. Furthermore, normative business ethics raises questions about how to engage with non-human

				entities.” (Tóth et al., 2022, p. 4)
		AI ethics (deductive)	• Design and regulation of AI technologies based on ethical principles (ethics of AI) to enforce AI’s ethical behavior (ethical AI) • Related to deontological ethics (design principles) • Related concepts: machine ethics	“To address AI ethics, one needs to consider the ethics of AI and how to build ethical AI. Ethics of AI studies the ethical principles, rules, guidelines, policies, and regulations that are related to AI. Ethical AI is an AI that performs and behaves ethically. One must recognize and understand the potential ethical and moral issues that may be caused by AI to formulate the necessary ethical principles, rules, guidelines, policies, and regulations for AI (i.e., Ethics of AI).”(Siau & Wang, 2020, p. 74)
		AI governance (deductive)	• Frameworks for standards in AI research, development, and application to ensure safe and fair usage of AI • Based on different governance structures/layers: societal AI governance (government regulation, law, etc.), industrial AI governance (independent oversight, code of conducts, etc.), organizational AI governance (leadership guidelines, internal reviews, etc.), software AI governance (reliable systems, explainable AI/XAI, etc.) • Related to deontological ethics	“Focusing on the shift from ethics to governance, it offers a bird-eye overview of the developments in AI governance, focusing on the comparison between ethical principles and binding rules for the governance of AI, and critically reviewing the latest regulatory developments.” (Koniakou, 2023, p. 71)
		AI perceptions (inductive)	• Ethical perceptions of humans in the context of AI • Underlying beliefs on the ethical intention of other agents (human or AI) • Related concepts: moral intentionality and ethical narratives	“...for the moral intention of reporting billing error, we conducted a regression analysis ... The impact of the perceptions of AI as human (vs. machine) on the moral intention was positive and significant...” (Giroux et al., 2022, p. 1034)
		AI adoption (inductive)	• Increased use of AI at work for actions and decisions • Related concepts: AI maturity, AI transformation, AI alignment, AI ambidexterity	“This option would not be prudent and would only amount to deferring a problem. A second approach, which is conservative, seeks to address the ethical issues only when they materialize or when they are reliably predictable. This approach would also not be prudent in the case of emerging technology, where the pace of adoption is significantly faster than other types of technology.” (Munoko et al., 2020b, p. 5)
		Human-Centered AI (HCAI) (inductive)	• Concerning to limit dangers and increase the benefits of AI for individuals, organizations, and society by putting the humans at the core of AI design, usage, and deployment • Based on ethical principles for HCAI • Related concepts: sustainable AI, AI divide, inclusiveness, equality	“Since HCAI systems are developed and implemented in multiple organizational structures, I propose 15 recommendations at three levels of governance: team, organization, and industry. The recommendations are intended to increase the reliability, safety, and trustworthiness of HCAI systems: (1) reliable systems based on sound software engineering practices, (2) safety culture through business management

				strategies, and (3) trustworthy certification by independent oversight. “ (Shneiderman, 2020, p. 26)
Presence of ethical theory	Identifies whether the ethical theory is explicitly mentioned or implicitly applied in the research.	Explicit (<i>deductive</i>)	An ethical theory is directly mentioned, named, or defined in the research	“...virtue ethics, understood as an approach to normative ethics that emphasizes moral character in contrast to approaches that emphasize duties and rules (deontology) or consequences of actions (consequentialism)” (Vanhée & Borit, 2022, p. 619)
		Implicit (<i>deductive</i>)	The research does not directly mention an ethical theory, but its principles are recognizable through the discussion, scenarios, or examples provided	“I can find no scientific proof that removing all our emotions would make us better people or even more effective in society, furthermore there is an inherent assumption that human biases and emotion are faults, but I would argue they are a rich part of our humanity.” (Seymour, 2018, p. 8)
Moral agency	An agents’ intention, beliefs, and reason to act morally or immorally based on its abilities to weight between different moral maxims and to critically reflect and make a free choice toward a more responsible option.	Human moral agency (HMA) (<i>deductive</i>)	- A human beings’ ability to make moral choices based on some notion of right and wrong and to be held accountable for these actions. - Related to virtue ethics - Similar concepts: traditional moral agency	“We tend to maintain human moral agency as superior because we believe it includes some intervening layers that the artificial moral agency might not ever achieve.” (Manna & Nath, 2021, p. 145)
		Artificial moral agency (AMA) (<i>deductive</i>)	- AI technologies which are designed to recognize unmorally requests and morally relevant aspects of a situation/action - There is an ongoing discussion on whether AI can have a moral agency to, at least, some extend (the opinions depend on the researchers’ narratives) - Related to deontological ethics	“While the definition of implicit ethical agent meets [the criteria] of artificial moral agent (AMA), the definition of explicit ethical agent meets the definition of artificial autonomous moral agent (AAMA).” (Serafimova, 2020, p. 2)
		Hybrid moral agency (<i>inductive</i>)	- All entities (stakeholders and technologies) that are somehow involved before, during, or after the human-AI interaction have some extent of moral agency (each entity a specific part of human agency) - Similar concepts: distributed morality	“As an AI’s action or decision can be traced back to a long, complex chain of interactions between human and the system—from developers to designers to users, each with different motivations, backgrounds, and knowledge...” (Sullivan & Fosso Wamba, 2022, p. 3)
Moral patients	Moral patients are entities who have limited autonomy and a limited ability to self-legislate their behavior due to age, physical, mental conditions,	Artificial moral patient (<i>inductive</i>)	- Technologies being based on AI are perceived as moral patients as they are affected by human beings and as their interactions are limited. - AI technologies can be moral proxies for human (moral) agents delegating certain moral interactions to them (moral agents are those who have moral obligations, and moral patients those to whom obligations are owed)	“I term psychological moral patency, that is, a form of moral status that may arise in virtue of a possession of specific psychological capacities such as sentience, autonomy, desires, and so on. I would suggest that this notion of psychological moral patency is of great relevance to much of our moral decisionmaking, insofar as capacities for undergoing suffering or exercising autonomy are relevant to many difficult decisions we make” (Shevlin, 2021, p. 460)
Ethical agents	Agents with the abilities to choose actions (based on their moral agency, see above) with ethical consequences,	Explicit ethical agents (<i>deductive</i>)	- Agent that acts on the behalf of ethical principles - Require cognitive intelligence (idea of narrow AI) - Related to deontological ethics	“Explicit ethical agents, on the other hand, are a tier higher than implicit ethical agents. This is because such machines already have some capacity to exhibit moral reasoning. At the top-most level of this hierarchy are full

	and thus being able to take responsible for those actions.			ethical agents. Average adult human beings are an instance of this type of agent.” (Boyles, 2018, p. 195)
		Implicit ethical agents (<i>deductive</i>)	• Agents those internal functions promote ethical behavior or avoid unethical behavior • Reasoning on ethics is based on the calculus of risk in each case (idea of being relevant for all kinds of autonomous systems) • Require logical reasoning • Related to consequentialism	“These machines are evaluated based on the moral consequences they produce. Next, machines that have built-in safety features could be considered as implicit ethical agents.” (Boyles, 2018, p. 195)
		Full ethical agents (<i>deductive</i>)	• Require consciousness, intentionality, and free will • Require multi-intelligence of human beings (idea of artificial general intelligence AGI or strong AI) • Related to virtue ethics	“A full ethical agent, on the other hand, ... would be similar to a robust explicit ethical agent, but it would also be even stronger in certain respects. According to Moor, it would be an agent that resembles an ‘average adult human’ in some ethically salient ways, namely, consciousness and free will.” (Cook, 2023, p. 163)
Moral responsibility	The extent to which an action is blameworthy, and an agent is, or several agents are ought to be accountable for violating standards of conduct by either behaving in an unacceptable manner or failing to behave in an acceptable manner.	Shared/distributed (moral) responsibility (<i>inductive</i>)	• The four requirements for moral responsibility (awareness of action, moral significance, consequences, and alternatives) are distributed between several agents • There is a scattered responsibility of several agents as actions are handled by different agents at the same time or at different time slots that are somehow interwoven • Related to moral agency • Similar concepts: responsibility attribution, accountability, legitimacy, judgment	“Whatever system of moral responsibility we use to ascribe responsibility to AI must cohere with the system that we already have. No matter how metaphysically different we feel that AI systems are from human agents, we must come up with a way to ascribe moral responsibility to them in a way that is intelligible and intuitive in the moral world that currently exists.” (Smith & Vickers, 2021, p. 485)
Moral intentionality	The underlying intention for an action or decision.	Moral intention (<i>inductive</i>)	• Intention of a person (inherent maxims to decide ethically) • Evaluated by others • Related to AI perceptions	“We show that moral intention (intention to report an error) is less likely to emerge for AI checkout and self-checkout machines compared with human checkout. In addition, moral intention decreases as people consider the machine less humanlike” (Giroux et al., 2022, p. 178:1027)
Ethics in human-AI interaction	Concerns the context in which ethics in human-AI interaction is researched and offers a categorization of ethical considerations on how both agents (AI and humans) can influence their ethical behavior reciprocally and how this is impacted by the environment in which the	Regulating human-AI interaction (<i>inductive</i>)	• Institutional, societal, or industrial regulation on AI (AI ethics, AI policies, AI acts) or humans (human rights, civil law, business ethics) • Related to deontological ethics	“Unfortunately AI advertising is not subjected to oversight, and so Developers can advertise unrealistic capabilities that are likely not backed up by evidence. Part of the solution may involve regulation, and one approach could be to study advertising regulation models like in medicine, and mandate that AI advertising similarly include a discussion of limitations, risks, and known weaknesses of the models. Specifically for affectively-aware AI, Regulators should mandate that advertising include a discussion of the demographics of the

	interaction takes place.			Emoters in the data that the models were trained on..." (Ong, 2021b, p. 5)
		Designing human-AI interaction (<i>inductive</i>)	• Design, training, evaluation, and development of AI technologies and guidance of those stakeholders involved (e.g., developers, software designers, data scientists) • Related to applied ethics	"As discussed earlier, the extinction of human beings may be a likely consequence if safety measures are not considered in designing AI systems. This is the reason why the research area of machine ethics looks into the possibility of modeling the first batch of AIs as being considerate of human values as well." (Boyles, 2018, p. 192)
		Managing human-AI interaction (<i>inductive</i>)	• Managerial approaches and decision-making on AI deployment, implementation, and involvement in strategic and operational processes (also human resource management, change management, knowledge and skill management) • Related to descriptive and applied ethics	"In order to respond to the ethical issues of operating AI solutions and manage the required practices of the model development, some interviewees felt that a centralized structure is great for identifying ethical concerns and taking action proactively." (Ghatar et al., 2023, p. 6)
		AI in education (<i>inductive</i>)	• Ethical training, teaching, and education in the context of AI • Related to applied ethics	"Before and after they ran the curriculum, we interviewed teachers to understand their opinions on pedagogical approaches to teaching AI, meeting students' needs, and the feasibility of doing the activities in the classroom. Our results indicate that, with appropriate training, even teachers who were new to computer science felt prepared and successfully engaged their students in the topic." (Williams et al., 2021, p. 15678)
		Theorizing/Researching upon human-AI interaction (<i>inductive</i>)	• Philosophical thoughts and theorizing on ethics in human-AI interaction • Related to meta-ethics	"Therefore, the main proposition of this text is that questions about the ethics of autonomous intelligent systems and AI platforms evolving over time through learning from data (Machine Ethics) cannot be answered by the concepts and thought experiments of the philosophy of mind and general ethics. These instruments are closed to the possibility of empirical falsification, use special sci-fi tools, are based on faulty analogies, transfer the burden of proof to the counterparty without justification, and usually end in an epistemological fiasco. Therefore, they do not bring any added value. Finally, let us stop analysing and overcoming these infertile philosophical distinctions and leave them at their own mercy." (Hauer, 2020, p. 426)
Research Method, Design,	Describes the underlying methodological approaches.	Syntheses (MacInnis, 2011) (<i>deductive</i>)	• Research method focusing on the identification of relevant material and the aggregation and classification/clustering of the material	e.g., Tasioulas (2018), van der Voort et al. (2015)

Approach, or Philosophy			- Summarizes explicit and tangible material (facts, principles, law, etc.) - Related terms: summary, typology, mapping studies	
	System Design (Siau & Rossi, 2011) <i>(deductive)</i>	- Design-oriented research, focusing on the development, testing, or validation of a system or data architecture, specific features, process modelling or software design - Related terms: design science	e.g., Arnold & Schuetz (2018)	
	Conceptualization (MacInnis, 2011) <i>(deductive)</i>	- Research Method focusing on the aggregation, restructuring, and interpretation of information (bringing material together in a novel order) - Aggregation of academic content - Related terms: framework, theory-building, theorizing	e.g., Vanh��e & Borit (2022), Shneiderman (2020)	
	Philosophical Essay (Creswell, 2014, pp. 5–11) <i>(deductive)</i>	- Philosophical worldviews - Summarizes implicit and intangible thoughts, opinions, and beliefs - Related terms: academic essay	e.g., Susser (2019), Swanepoel (2021)	
	Position Paper (Creswell, 2014, pp. 5–11) <i>(deductive)</i>	- Research that represents the aggregation of an own opinion and the process of making claims - Evidence is established by theoretical justification and the streamline of argumentation	e.g., Unver (2023)	
	Case Study (Creswell, 2014, p. 13) <i>(deductive)</i>	- Deep-dive investigation of one or more single companies - Besides interviews, the company’s material is analyzed in depth	e.g., Vakkuri et al. (2019), Williams et al. (2021)	
	Experiment (Creswell, 2014, pp. 13, 170–174) <i>(deductive)</i>	- Research that seeks to determine a specific influence by using specific treatments - Performing a scientific test in which participants have to perform a series of actions and researchers carefully observe their effects in order to learn about something - Can be online or offline	e.g., Willems et al. (2022), Vanderlest & Winfield (2018)	
	Quantitative Online Survey (Creswell, 2014, pp. 155–173) <i>(deductive)</i>	- Based on an online questionnaire with measures and scales - Statistical or descriptive analysis	e.g., Khan et al. (2022)	
	Quantitative Text Analysis (Creswell, 2014, pp. 46–47) <i>(deductive)</i>	- Analyzing textual content from various sources (e.g., social media, documents) - Statistical or descriptive analysis	e.g., Neubert & Montanez (2020)	
	Qualitative Content Analysis (Creswell, 2014, pp. 186, 197–201) <i>(deductive)</i>	- Identifying core consistencies and meanings through the systematic classification process of coding and identifying themes or patterns of texts - Related terms: grounded theory, explorative investigation, deductive and inductive coding	e.g., Polyviou & Zamani (2022)	
	Interviews (Creswell, 2014, p. 190) <i>(deductive)</i>	- Question- and answer situation with a researcher and research subject on a specific topic based on an interview guideline - Synchronic dialogue of at least one interviewee and at least one interviewer - Can be online or offline or via phone - Related terms: semi-structured interviews	e.g., McDonald & Pan (2020), Stahl et al. (2022)	

Appendix VIII. Bibliometric Analysis

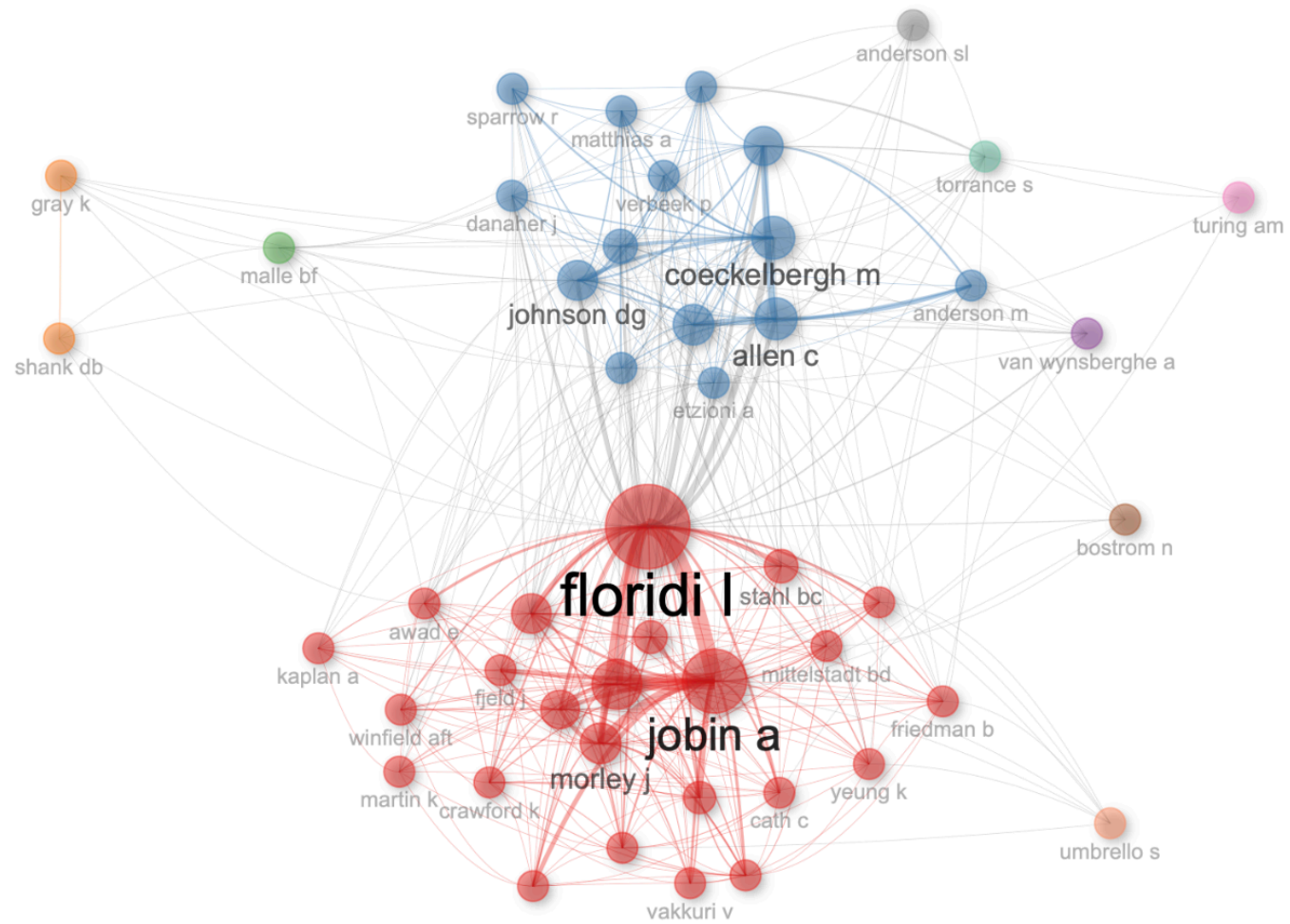
The 14 articles in Table 11 for which metadata could not be retrieved were absent from the Dimensions.ai database, a comprehensive open-source alternative to Web of Science (WoS), and thus, were not used for bibliometric analysis. The primary rea-sons for their omission include their nature as conference proceedings and the lack of DOIs (Digital Object Identifiers).

Table 11 Excluded articles from the bibliometric analysis

Authors	Year	Title	Outlet	J/C/Others	Type
Bilal et al.	(2021)	Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics	ACIS	Conference	Proceedings
Eicher et al.	(2018)	Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies	AIES	Conference	Proceedings
Jantunen et al.	(2021)	Building a Maturity Model for Developing Ethically Aligned AI Systems	IRIS	Conference	Selected Seminar Papers
Kučera	(2019)	Artificial Intelligence in the Context of Managerial Responsibility	IMES	Conference	Proceedings
Mäntymäki et al.	(2023)	Designing an AI governance framework: From research-based premises to meta-requirements	ECIS	Conference	Proceedings
Mayer et al.	(2020)	How corporations encourage the implementation of AI ethics	ECIS	Conference	Proceedings
Minkkinen & Mäntymäki	(2023)	The Institutional Logics Underpinning Organizational AI Governance Practices	SCIS	Conference	Proceedings
Polyviou & Zamani	(2022)	Towards Europe's AI Strategy: Desires & Realities	ECIS	Conference	Proceedings
Schneider et al.	(2023)	How individuals can shape AI through data - An AI literacy and morality perspective	ECIS	Conference	Proceedings
Seppälä et al.	(2021)	From Ethical AI Principles to Governed AI	ICIS	Conference	Proceedings
Susser	(2019)	Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures	AIES	Conference	Proceedings
Vidu et al.	(2020)	Artificial Intelligence and the Ethical Use of Knowledge	STRATEGICA	Conference	Proceedings
Walke et al.	(2023)	AI in Government: A Study on Explainability of High-Risk AI-Systems in Law Enforcement & Police Service	WI	Conference	Proceedings
Westerstrand	(2023)	Ethics in the intersection of AI and democracy: The AIDEM Framework	ECIS	Conference	Proceedings

Appendix IX. Author Citation Network

Fig. 6 Zoom-in in the author citation network (n=50 top cited authors)



Appendix X. Distribution of Articles over Years

Fig. 7 Average citations of the citation network per year (n=137)

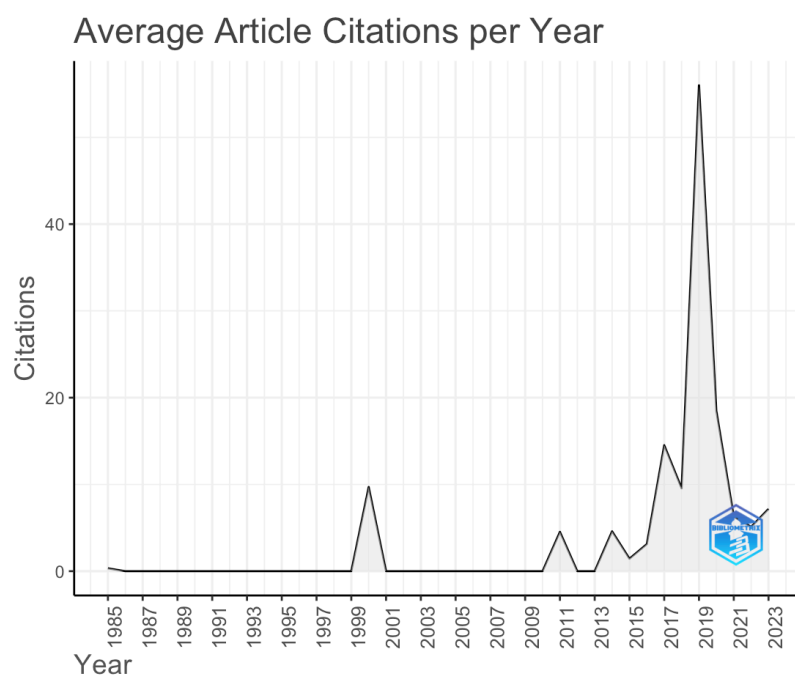
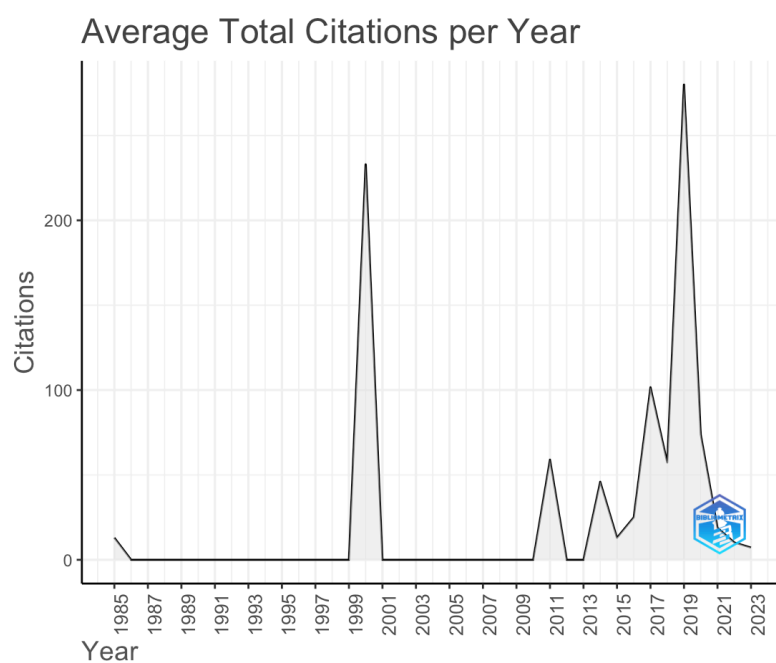


Fig. 8 Total citations of the citation network per year (n=137)



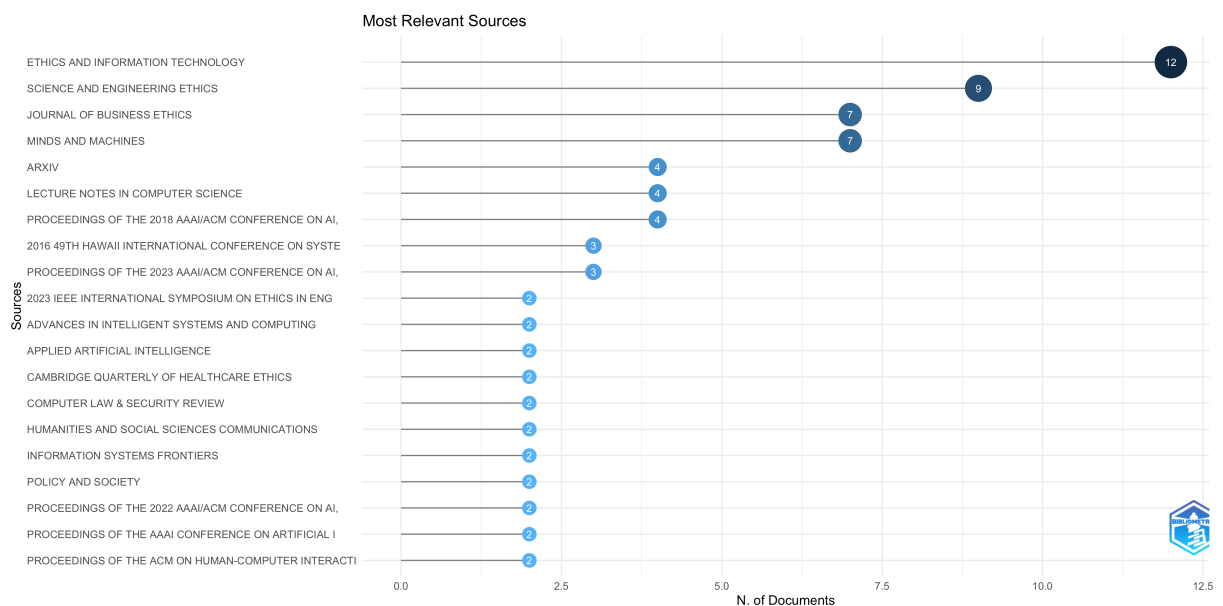
Appendix XI. Most Cited Publication Outlets

We further employed bibliographic coupling analysis (node size= 20). We analyzed the level of articles as seen in Figure 9 using the Walktrap clustering algorithm.

Fig. 9 Most cited publication outlets (n=20)



Fig. 10 Most relevant sources (n=137)



Appendix XII. Author Citations based on Authors Countries

Fig. 11 Country distribution of authors (n=6,523)

Country Scientific Production

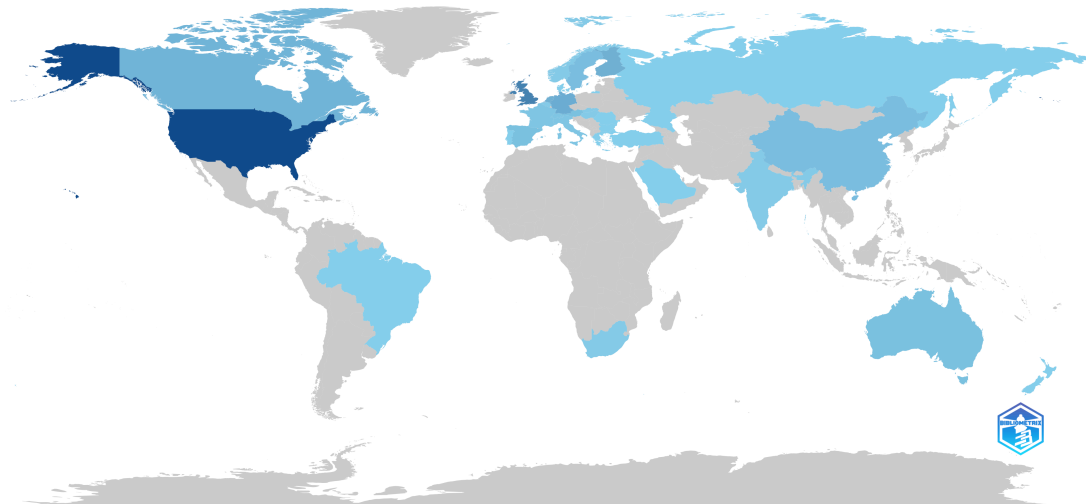
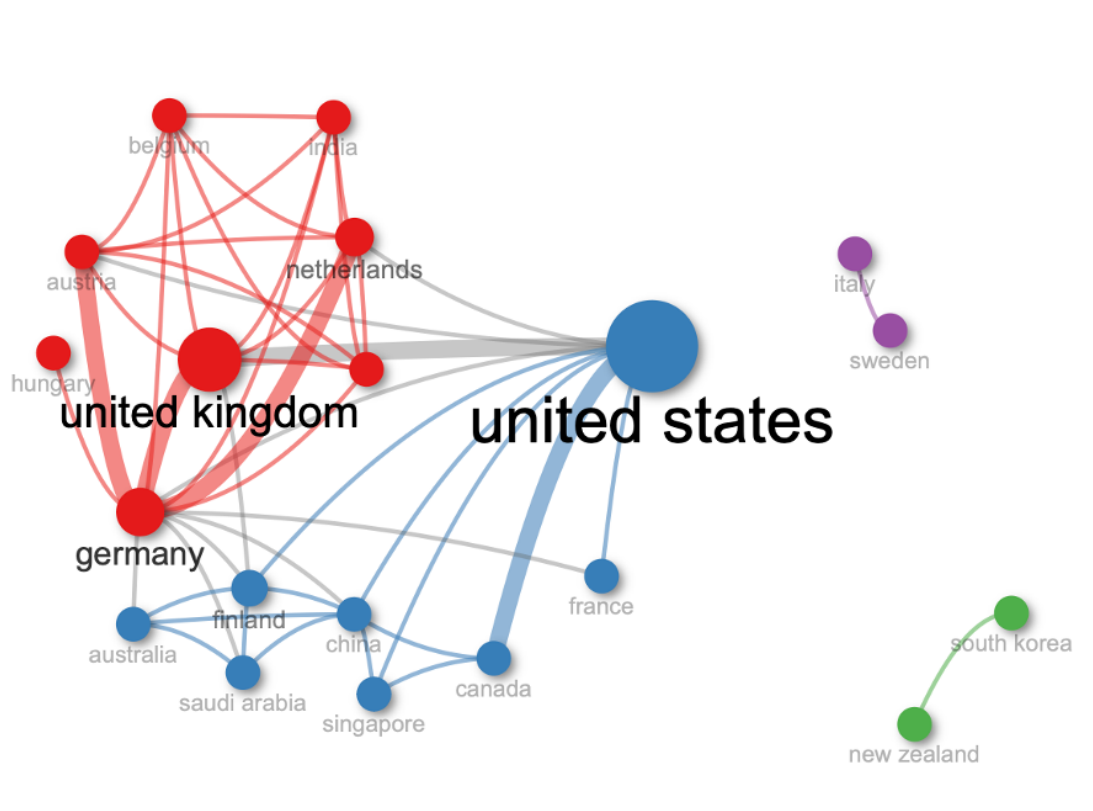


Fig. 12 Collaboration network of the final sample's first authors (n=137)



Appendix XIII. Perceptions of HMA compared to AMA

Table 12 Key findings on perceptions of HMA and AMA

Different perceptions of HMA compared to AMA	Percentage	N
AI is perceived as an explicit ethical agent having moral agency through programmed principles.	4.64 %	7
AI is perceived as an implicit ethical agent having moral agency through unsupervised learning.	3.97 %	6
Humans are perceived as the only full ethical agents so that AMA requires moral updates through human.	3.31 %	5
AI is perceived as lacking moral agency so that responsibilities remain with humans.	9.93 %	15
HMA is perceived as guiding AMA strengthening the importance of aligning AMA with human values.	12.58 %	19
AMA is perceived as opaque which can lead to moral uncertainties among humans potentially limiting HMA.	9.27 %	14
AMA is perceived as strengthening HMA implying that both are necessary for ethical human-AI interaction.	13.25 %	20
The interdependencies of HMA and AMA are perceived to require a shared responsibility among human stakeholders.	31.79 %	48
HMA is perceived as being strengthened through training and discussion.	5.96 %	9
HMA is perceived to differ from AMA as they have differences in emotionality and moral reasoning.	5.30 %	8

Appendix XIV. Taxonomy of Mechanisms Fostering Morality Within Human-AI Interaction

Table 13 Taxonomy of mechanisms that can be used to foster morality within human-AI interaction

Mechanism	Driving entities	Ethical Theory	Authors	Research Gaps in the Context of Descriptive ethics
Ethical principles guiding AI design impacting the underlying processing model of the AI drive ethical behavior of the AI.	AI developers of AI that establish and follow ethical principles in the design of AI technologies (<i>the AI is directly influenced by the developer's environment</i>). This can further lead to an indirect influence of the human user, when the AI technology is designed to actively avoid misuse (<i>the human user is indirectly influenced by the AI's design</i>).	Deontology	e.g., Boyles (2018), G�nova et al. (2023), Hauer (2020), Manna & Nath (2021), Martinho et al. (2021)	There remains a gap in the exploration of ontological perspectives on ethics at a meta-level within the context of AI development. Specifically, there is a need for research investigating the underlying philosophical assumptions and implications surrounding the development of AI technologies designed to imitate ethical behavior, which could contribute to a deeper understanding of the ethical implications and consequences of such technological advancements.
CSR principles guiding AI deployment in organizations impact how humans create data input for the AI and work with the data output of the AI, thus, influence ethical behavior of employees as human users.	Managers establish guidelines for AI usage in organizations to reduce risks and increase business value (<i>the human user is influenced by the organizational environment</i>). There can be carrot and stick mechanism that enforce norm-compliance.	Consequentialism	e.g., Bankins & Formosa (2023), Millar (2016), Pruski (2023), T�th et al. (2022), Swanepoel (2021)	There is a lack of investigation into how different contextual factors shape the operationalization of moral proxies in determining the ethical responsibilities attributed to AI systems within distinct sectors, such as military, healthcare, or marketing.
Norms and policies imposed upon AI technologies per law or contracts define how to deal with the data input that the AI requires and how the AI can create what kind of data outputs	Governmental or AI initiative representatives and politicians that define law and policies guiding what AI technologies are allowed to do and not (<i>the AI is influenced by the regulatory environment</i>). As the AI technology itself cannot be judged prosecution or carrot and stick mechanisms are directed at AI providers.	Contractarianism	e.g., G�mez de �greda (2020), Niederman & Baker (2023)	There is a need to investigate how differences in social norms and regulatory environments, exemplified by varying AI regulations in the USA, China, and Europe, impact the feasibility and acceptance of universal ethical standards for AI development and deployment. Bridging this gap would provide valuable insights into the challenges and opportunities associated with harmonizing ethical principles in the context of AI governance on a global level.
Teaching ethical norms combined with AI knowledge within AI education enhances human users' competencies, skills and knowledge (AI	Educators, trainers, professors, teachers, and so on that help human users to develop abilities improving their interactions with AI technologies in a successful and responsible way (<i>the</i>	Virtue Ethics	e.g., Raji et al. (2021), Schneider et al. (2023), Williams et al. (2021)	There is a need for empirical case studies examining the implementation and effectiveness of AI ethics courses. By conducting such case studies, researchers can gain insights into the practical challenges and opportunities associated with integrating ethical education into AI curricula.

literacy) and impact their abilities to create and evaluate data input for the AI and to critically reflect and successfully use data output of the AI.	<i>human user is influenced by the societal environment).</i>			
Ethical behavior or human users' corrections on AI's output impact the AI's processing models since (un)supervised learning mechanism of the AI follow human data input and feedback.	Humans can help to increase ethical behavior of the AI by integrating ethical knowledge into the AI's data base and providing feedback on unmoral behavior to <i>(human user's feedback influences the AI's underlying processing models and outputs).</i>		e.g., Etzioni & Etzioni (2017), Greene et al. (2019), Roberts (2023)	Despite the acknowledgment of the potential for human feedback to influence the ethical behavior of AI systems, there remains a significant gap in understanding the practical implementation and effectiveness of this mechanism. Specifically, there is a lack of empirical research examining how human feedback, particularly regarding unethical behavior, can impact the underlying processing models and outputs of AI systems.
Collective deliberation and sense-making on several areas of AI design, regulation, deployment, education, and so on help to provide a more diverse discourse on how, when, why, and under which conditions ethics should be integrated into AI technologies (thus, impacting the subject areas and tasks of AI) by considering different moral compass of humans from different societal and regulatory environments.	International ethical committees, ethical boards, chief ethical officers, and so on, can define a global baseline of AI governance <i>(members of those ethical discourses influence other deciders upon AI technologies implementation, education, usage, and so on; thus, this demonstrates an influence within the environment of human-AI interaction).</i>	Applied ethics	e.g. Etzioni & Etzioni (2017), Dodig Crnkovic & Çürüklü (2012)	Despite the establishment of international ethical committees and boards to govern AI at a global level, there is limited empirical evidence on the comparative effectiveness of these mechanisms in shaping AI governance and decision-making processes. Furthermore, there is a lack of research examining how the diverse moral compasses of human users influence the ethical development and deployment of AI technologies.
Considerations on whether we require more context-specific or other ontological views on how we perceive the world in an ethical sense fosters a holistic reflection upon ethics.	Theorizing and philosophizing on ethics and AI in general and the narratives humans show regarding both constructs can drive ethical behavior or not <i>(human perceptions influence the AI).</i>	Meta-ethics	e.g., Siapka (2022)	More theoretical work is required on how cultural differences in ontological worldviews influence interpretations, relevance, and prioritization of ethical theories.

Paper II

Hafner, A.; Hammerschmidt, T. & Stolz, K. (2024). *Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review*. Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany, <https://aisel.aisnet.org/wi2024/46>.

Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review

Research Paper – Accepted Version

Alina Hafner¹, Teresa Hammerschmidt², Katharina Stolz³

¹ Technical University of Munich, TUM School of Management, Heilbronn, Germany
alina.hafner@tum.de

² University of Bamberg, Information Systems and Social Networks, Bamberg, Germany
teresa.heyder@uni-bamberg.de

³ University of Stuttgart, Innovation and Service Management, Stuttgart, Germany
katharina.stolz@bwi-uni-stuttgart.de

Abstract. Artificial Intelligence (AI) ethics research is a multifaceted field, requiring different theoretical justifications in which researchers can ground their underlying perspectives on ethics. We provide an overview of the major normative ethical theories used in Information Systems research on AI ethics. Through a systematic scoping review, we assess the prevailing theories, their progress, and areas needing further study. Our findings reveal a dominance of deontological ethics, which results in determining ethics mainly from the AI’s perspective by discussing ethical design principles but not from how a human user’s virtue ethics perspective guides humans’ moral behavior when collaborating with AI equally. We suggest that researchers recognize how normative ethical theories might bind their work, impacting their understanding of moral agency and responsibility and guiding Corporate Digital Responsibility practices for organizations striving for responsible AI design, deployment, and usage.

Keywords: Scoping review, AI ethics, ethical theories, digital responsibility

1. Introduction

With Artificial Intelligence (AI) able to the ability “to perform cognitive functions that we associate with human minds” (Rai et al., 2019, p. iii), a novel generation of technologies arises, exhibiting greater autonomy and complexity compared to previous ones (Baird & Maruping, 2021; Brendel et al., 2021). When AI informs, controls, or even automates specific tasks and decisions, questions of morality and responsibility arise – especially when unintended outcomes, such as the discrimination of certain groups of people due to data biases, occur (Giermindl et al., 2022; Kellogg et al., 2020). Novel forms of human-AI collaboration present several ethical challenges for individuals, organizations, and society (Dwivedi et al., 2023). This leads to a rising academic discourse in the context of Corporate Digital Responsibility (CDR) concerning responsibility questions that arise with the widespread use of autonomous technologies (Mihale-Wilson et al., 2022) to form a transformation that is “not only ...

driven by what is technologically possible, but also what is socially desirable and sustainable.” (Mueller, 2022, p. 689). CDR and AI ethics are interconnected as CDR encompasses the ethical use of digital technologies, including AI, ensuring that AI systems are designed, deployed, and utilized in ways that align with ethical principles and promote societal well-being (Sullivan & Wamba, 2022; Tóth et al., 2022).

To contribute to CDR research, we must be aware that different normative thoughts on ethics impact researchers’ understanding of AI ethics and, thus, how they define responsible AI design, deployment, and usage (Lobschat et al., 2021). Siau and Wang (2020) stated that research on AI ethics concerns designing AI technologies according to ethical principles (ethics of AI) to enforce AI’s ethical behavior (ethical AI). The perspective that AI must follow defined rules draws inspiration from deontological ethics as one normative ethical theory used by Weber (1988) and Kant (1797) and informs research to design AI in an ethically responsible manner based on duty-bound principles (e.g., Floridi et al., 2018; Jobin et al., 2019). However, AI ethics also considers how users’ virtue ethics – representing humans’ inherent moral understanding of right and wrong – impact ethical behavior and actions when interacting with AI (Heyder et al., 2023; Stolz et al., 2024). Researchers must be aware of how different ethical theories impact the understanding of responsible AI when transferring academic insights into practices (Brendel et al., 2021; Dignum, 2019).

The underlying ethical theory also influences how researchers perceive the moral agency of humans compared to AI (Coeckelbergh, 2020) – mainly when discussing who is responsible for undesired outcomes (Mueller, 2022). For instance, Allen et al. (2000, p. 252) argued that there is a “deep disagreement in ethical theory [...] about what standards moral agents ought to follow.” Those who follow the principle of utility – believing actions should achieve the maximum of aggregated good consequences (consequentialism)– argue that autonomous technologies should be programmed to reach this maximum from the bottom-up (Allen et al., 2000; Brendel et al., 2021), while others strive for bottom-up modeling of AI technologies based on virtue ethics theory making the “agents of AI (users, developers) ... able to take responsibility for what they do with AI,” by arguing that humans have to remain responsible since AI does not meet the traditional criteria of moral agency (Coeckelbergh, 2020, p. 2062).

Thus, different normative ethical theories impact our understanding of moral agency and AI responsibility from different perspectives (Dignum, 2019). To understand the contribution of AI ethics research and derive suggestions for responsible AI design, deployment, and usage following CDR (Mihale-Wilson et al., 2022), knowing which ethical theories ground the theoretical framing of AI ethics is essential. We overview the major normative ethical theories used in Information Systems (IS) research on AI ethics by focusing on deontological ethics, consequentialism, and virtue ethics (Bankins & Formosa, 2023; Bilal et al., 2021). Our research questions (RQs) are:

(RQ1): *What normative ethical theories are commonly used in AI ethics research?*

(RQ2): *What topics of AI ethics are studied the most within the IS field?*

(RQ3): *What are the calls for future investigations on AI ethics research?*

To answer these questions, we conduct a scoping review as it is appropriate to summarize academic discourse on a specific topic to identify the existing research landscape and potential gaps (Paré et al., 2015). We systematically review research on

AI ethics within the IS field to determine which normative ethical theories focus on existing IS research and highlight areas that require further investigation.

Our contributions are twofold: *First*, we demonstrate how different normative ethical theories, implicitly or explicitly as theoretical frameworks, influence the interpretation of results and the discussion of findings in AI ethics research. The latter element is particularly relevant given the strong inclination towards deontological ethics, especially within the IS field, guiding researchers' suggestions for responsible AI design, deployment, and usage by combining ethics and human-computer interaction and technology application research as two disciplines of CDR (Lobschat et al., 2021). *Second*, we discuss the most studied topics in AI ethics research and offer recommendations for future research and practitioners to maximize the benefits of AI transformation.

2. Theoretical Background and Related Work

2.1. Normative Ethical Theories

Ethical theories guide how humans evaluate and form moral judgments (Beauchamp & Bowie, 1988). Within the domain of ethical studies, besides normative ethics, descriptive ethics, metaethics, and applied ethics can be analyzed (Brendel et al., 2021). We focus on normative ethics as it offers a systematic approach to determining moral duties and ethical conduct, providing prescriptive principles that dictate what actions are morally right or wrong (Brendel et al., 2021). Normative ethics provide guidance on how people should act. It encompasses various ethical theories, each proposing different principles for moral actions (Gewirth, 1960; Weaver & Trevino, 1994). Figure 1 provides essential normative ethical theories following Bankins and Formosa (2023):

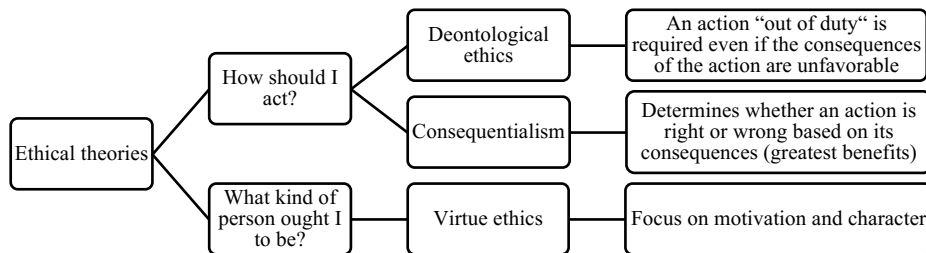


Figure 3. Normative Ethical Theories

Deontological ethics roots itself in duty and obligation, positing that specific actions are inherently moral or immoral, regardless of their outcomes (Gaus, 2001). Immanuel Kant introduced the concept of the “categorical imperative”, which asserts that an action is morally right if it can be universalized without contradiction (Kant, 1797).

Consequentialism asserts that the morality of an action depends solely on its outcomes or consequences, determining the overall goodness or badness of results (Yu et al., 2018). The most well-known variant of consequentialism is utilitarianism, but other forms exist (e.g., ethical egoism), each with a different perspective on what

outcomes are to be valued (Bentham, 2007; Mill, 2009). We focus on utilitarianism as it posits that the morality of an action is determined by the extent to which it maximizes overall happiness or pleasure (Bentham, 2007; Mill, 2009).

Virtue ethics, based on the works of Aristotle (Brown & Aristotle, 2009), emphasizes the importance of cultivating virtuous character traits, such as courage and honesty, and posits that moral actions stem from the inherent moral values of individuals (Eitel-Porter, 2021; Heyder et al., 2023). Vallor (2016) claims that, as emerging technologies grow more complex, virtue ethics provides a framework for moral decision-making amidst uncertainty, guiding us toward living well with these advancements.

2.2 Understanding of Moral Agency Forming Responsibility Perceptions

Central to normative ethical theories is the concept of moral agency that determines our perceptions of responsibility (Coeckelbergh, 2020; Passlack et al., 2023). Moral agency encompasses prevailing notions of “right” and “wrong” within a particular society and time, defined by a set of established norms and rules, both explicit and implicit (Behdadi & Munthe, 2020). It demonstrates the ability to choose the morally more responsible option (Champagne & Tonkens, 2023; Søvik, 2022). These moral values, norms, and belief systems constantly evolve and intertwine with historical social processes. Various social entities, such as religious institutions, criminal organizations, or cultural groups, craft their unique moral codes and adapt them over time, shaping distinct moralities. Hence, moral agency is required to attribute responsibility for specific actions (Constantinescu et al., 2021).

Central to moral agency is the concept of inner free will, which means the capacity to choose between actions, whereas AI does not have a moral agency in a traditional sense (Coeckelbergh, 2020). However, the increased autonomy of AI might lead people to perceive an Artificial Moral Agency (AMA) similar to a traditional human one and attribute responsibility to the AI, resulting in responsibility gaps (Behdadi & Munthe, 2020; Coeckelbergh, 2020; Giermindl et al., 2022). Thus, the CDR discourse rises (Lobschat et al., 2021) concerning ethical tensions from different responsibility perceptions of AI design, implementation, and usage.

Lobschat et al. (2021, p. 875) define CDR as “the set of shared values and norms guiding an organization’s operations with respect to four main processes related to digital technology and data.” While CDR focuses on the organizations’ value system (corresponding to virtue ethics) and norms (corresponding to deontological ethics) in forming responsible or non-responsible actions, we aim to go one step before the organizations’ CDR culture by investigating the researchers’ perspectives that might influence CDR behavior as social context (Lobschat et al., 2021) by providing suggestions for responsible AI design, deployment, and usage that aligns human values and considers the ethical impact of AI based on different ethical theories (Dignum, 2019, p. 47).

2.3 Existing Literature Reviews on AI Ethics

Several systematic literature reviews summarize existing research on AI ethics within the IS field, each contributing unique insights and identifying prevailing ethical

principles (Berente et al., 2021; Mirbabaie et al., 2022). However, a nuanced gap remains in understanding precisely which ethical theories have been used to better explain the underlying motivations for their use and its consequences on theoretical and practical contributions made by AI ethics researchers. For instance, Floridi et al. (2018) laid a significant foundation by introducing critical ethical principles for the establishment of a “Good AI Society” such as beneficence, non-maleficence, autonomy, justice/fairness, and accountability (p. 689). These principles ensure that organizations prevent technology from harming humankind. Jobin et al. (2019) expanded the scope by including grey literature in their systematic scoping review. They identified five primary ethical principles (transparency, justice and fairness, non-maleficence, responsibility, and privacy) and highlighted additional ones, broadening AI’s moral discussion. Vainio-Pekka et al. (2023) further investigated the use of ethical principles specifically for explainable AI, reiterating the importance of the primary principles identified by previous research and pointing out the vagueness and lack of ethical knowledge as significant challenges. Khan et al. (2021) focused on ethical challenges and principles, echoing concerns about the vague nature of ethical guidelines and the prevalent lack of ethical knowledge. As Siau and Wang (2020) argued, research on AI ethics should not solely focus on exploring ethical principles for AI design but also on how to foster ethical behavior of the AI.

These reviews collectively emphasize the critical role of ethical principles in AI development, often referencing Jobin et al.’s (2019) global landscape of AI ethics guidelines and Floridi et al.’s (2018) AI4people principles as pivotal in motivating further research. The existing literature underscores the importance of these ethical principles in guiding AI development but often needs to delve deeper into the moral philosophy guiding their perspective on AI ethics. This gap highlights a crucial area for future research. Filling this gap could significantly advance the findings of reviews like those conducted by Brendel et al. (2021) and Heyder et al. (2023), which aim to understand the management of human-AI interaction from an ethical perspective by using different normative ethical theories for their literature analysis. Both works highlight how different ethical theories impact our understanding of moral or immoral behavior in the context of aiming to achieve ethical AI design, usage, and deployment.

8 Method

Our methodology builds on a systematic scoping review. A scoping review is appropriate to answer our research questions as it provides an overview of the nature of research on a specific topic to identify the breadth of coverage and potential research gaps (Paré et al., 2015). Unlike a systematic review, which synthesizes high-quality evidence to answer specific RQs and assess study quality, a scoping review maps all relevant literature irrespective of study design (Arksey & O’Malley, 2005). The systematic nature of following the steps of Templier and Paré (2018) increases the transparency:

1. Search strategy: Given our focus on IS research, we searched the following four databases: Association for Computing Machinery Digital Library (ACM), Association

for Information Systems Electronic Library (AIS), Business Source Complete (BSC), and Institute of Electrical and Electronics Engineers Explore (IEEE). We searched for relevant literature by querying the title, abstracts, and keywords – if available – using the following search string: (“ethics” OR “morality” OR “ethical theory” OR “ethical family” OR “moral theory” OR “ethical philosophy” OR “moral philosophy” OR “normative ethics” OR “virtue ethics” OR “utilitarianism” OR “consequentialism” OR “deontology”) AND (“Artificial Intelligence” OR “human-AI interaction”). We decided on these keywords after conducting initial exploratory research. The search conducted in September 2023 yielded 1234 overall hits. After eliminating 75 duplicates, we yielded 1159 unique hits. Table 1 illustrates the search process:

Table 1. Snapshot of Coding

Steps	ACM	AIS	BSC	IEEE	Total
Literature search	141	229	336	528	1234
Duplicates	1	1	46	27	75
Unique hits	140	228	290	501	1159
Format Screening	424 excluded and 735 remaining articles				
Non-scientific	32	1	218	85	336
Non-English	1	0	20	0	21
Incomplete	11	16	1	1	39
Stand-alone review	6	13	4	5	28
Relevance Screening	674 excluded and 61 remaining articles				
No ethical theory	10	37	5	47	99
Missing own perspective	9	26	1	7	43
No AI technology	0	0	1	13	14
No human-AI interaction	37	88	27	271	423
No dedicated contributions	7	26	6	56	95
Final sample	27	21	7	6	61
Top IS	0	15	1	0	16

2. Literature screening: We applied format criteria through format screening, eliminating non-scientific literature (e.g., poster presentations, thesis), non-English articles, incomplete articles (e.g., research-in-progress articles, extended abstract), and systematic and standalone literature reviews. Given the underlying goals of a scoping review, we did not restrict our search to a particular time range or specific quality criteria (e.g., rankings, peer-reviewed literature). This resulted in the elimination of 424 articles, and 735 articles remained for relevance screening. For relevance screening, we applied relevance criteria. We excluded articles that: i) cannot be related to ethical theories given their focus on other theoretical lenses (e.g., theory of planned behavior), ii) do not offer a meta-perspective on ethics from the paper’s standpoint of view (e.g., ethics only appears in the context of an interviewees statement), iii) do not involve technologies that can be aligned to AI following the underlying definition of (Rai et al., 2019, p. iv) do not provide dedicated contributions regarding ethical theories (e.g., by solely focusing on one principle), do not focus on AI being used in several forms of human-AI interaction. 674 articles were excluded, and 61 articles remained as the final sample.

3. Classification of our final sample: Scoping reviews aim to be as comprehensive as possible in identifying primary studies and reviews (Arksey & O’Malley, 2005) and to achieve this, articles from journals of varying quality should be considered. To

address the differences in quality between the articles of the final sample belonging to the core of IS' top journals and conferences, we followed a systematic approach to categorize the final sample into "top IS" (n=16) and other articles of the final sample not belonging to this category (n=45). Our list of "top IS" journals and conferences is based on several sources²: we involved IS conferences and journals being part of the VHB JOURQUAL3 Ranking, the Financial Times Top 50 journal list ("FT50"), the "basket of the 11" following Lowry et al. (2013), the list of "pure IS" having a rating of 0.8 or higher as outlined by Walstrom and Hardgrave (2001), and the framework of assimilating journals being ranked A+, A, or B as proposed by Levy and Ellis (2006).

4. Qualitative analysis: The 61 articles of our final sample were analyzed through manual coding. First, we structured the samples' meta-information (e.g., regarding the research domain, method, and the ranking of the publication outlet) to provide a descriptive overview of the different article types within the final sample. We then opted for qualitative analysis to structure the content of relevant articles. We used the theoretical background to structure the findings mainly deductively (Mayring, 2014, pp. 95-98). To increase the quality of the coding process and redefine the coding schema, we used 32 articles for parallel training purposes of the coders based on the coding schema given in Table 2. Utilizing Cohen's Kappa, as Cohen (1960) outlined, we measured interrater agreement, obtaining scores of $\alpha=0.76$ in the initial round and $\alpha=0.82$ in the subsequent round. These results signify a substantial agreement between coders, which aligns with McHugh's interpretation (2012). To enhance coding reliability, we conducted multiple discussion rounds among coders. Table 1 presents a snapshot of the coding results; Table 2 shows the final sample related to different AI research areas.

9 Results

Regarding our **first research questions** on what normative ethical theories are most commonly used in AI ethics research, our review highlights a clear focus on deontological ethics, with 37.70% involving specific contracts to be obtained when designing or managing AI and 20.97% following specific ethical AI principles. Only two articles discussed all three common views on normative ethics: Allen et al. (2000) stated that understanding artificial moral agents varies with ethical theories. Vanhée and Borit (2022, p. 623) argued that using "classic ethical theories, such as deontic, utilitarian, and virtue ethics" as moral exemplars can impact ethical behavior. Hence, our final sample does not present virtue ethics and consequentialism (each 4.84%), demonstrating a research gap, especially as "moral values need to guide technologists in their research and to help decision-makers regulate the use that is made of the possibilities that AI presents from their deeper understanding of the possibilities and the consequences" (Gómez de Ágreda, 2020, pp. 3-4).

² <https://ethicaltheories.github.io/>

Table 2. Snapshot of Coding

Category	Explanation	Codes	Coding Rules	Examples
Normative ethical theory	States whose view on normative ethics is specified or assigned.	Deontology Consequentialism Virtue ethics Contractarianism Other	We use the definitions provided in Section 2. If the theory is not explicitly stated but discussed on a meta-level that aligns with a particular theory, it may be coded “implicit”; see category “presence of ethical theory”).	Code for virtue ethics: “Virtue ethics focuses on moral character development rather than duties, rules, or the consequences of actions to decide what is right or wrong.” (Bilal et al., 2021, p. 2) Codes for deontology & consequentialism: “The two approaches considered are utilitarianism, on the one hand, and Kant’s use of the ‘categorical imperative’ on the other.” (Allen et al., 2000, p. 252)
Presence of ethical theory	Identifies whether the ethical theory is explicitly mentioned or implicitly applied in the research	Explicit Implicit Absent	If the research does not directly mention an ethical theory but its principles are recognizable through the discussion, scenarios, or examples provided, we code it as “implicit.” If an ethical theory is directly mentioned, named, or defined in the research, we code it as “explicit.”	Code for explicit (virtue): “...virtue ethics, understood as an approach to normative ethics that emphasizes moral character in contrast to approaches that emphasize duties and rules (deontology) or consequences of actions consequentialism)” (Vanhée & Borit, 2022, p. 619) Code for implicit (deontological): “We are assuming that ethical and moral reasoning by humans is fundamentally computational and does not involve some special human-only force or quality.” (Seo & Thorson, 2022, p. 1077)

Table 3. Overview of the final sample related to different research categories

Category	Explanation	Example	All (62)	Top IS* (16)
Regulating AI ethics	Articles in this category have an institutional or regulatory focus on AI ethics.	Almeida et al., 2020*; Auld et al., 2022; Colmenarejo et al., 2022; Cooper et al., 2022; Fabiano, 2019; Ha, 2022; Henderson, 2019; Lim & Kwon, 2021; Polyviou & Zamani, 2022*; Seppälä et al., 2021*; Tidjon & Khomh, 2022; Unver, 2023; Walke et al., 2023; Weber, 2020; Westerstrand, 2023*; Wright, 2020	16	4
Designing AI ethics	Articles in this category mainly deal with the design of ethical AI, i.e., technical or design-oriented guidelines, as well as actual technical solutions or computational models.	Benner et al., 2021*; Bilal et al., 2021; Chaput et al., 2021; Cook, 2023; Eicher et al., 2018; Emdad et al., 2023*; Gerdes, 2022; Hooker & Kim, 2018; Jantunen et al., 2021; Ong, 2021; Schlimbach & Khosrawi-Rad, 2022*; Susser, 2019; Vakkuri et al., 2021; Vanderelst & Winfield, 2018; Vanhée & Borit, 2022	15	3
Researching AI ethics	Articles in this category contribute to research on AI ethics.	Aslan et al., 2022*; Dyrkolbotn et al., 2018; Estrada, 2018; Greene et al., 2019*; Hawkins & Mittelstadt, 2023; Mirbabaie et al., 2022*; Seymour, 2018; Siapka, 2022; Siau & Wang, 2020; Vainio-Pekka et al., 2023; Zhang et al., 2021	11	3
Managing AI ethics	Articles in this category mainly contain managerial approaches to dealing with AI ethics.	Agbese et al., 2023; Akbari Ghatat et al., 2023*; Corvite et al., 2023; Figueras et al., 2022; Gómez de Ágreda, 2020; Mäntymäki et al., 2023*; Mayer et al., 2021*; Minkkinen & Mäntymäki, 2023; Rismani & Moon, 2023; Wang et al., 2020*	10	4
Educating AI ethics	Articles in this category mainly introduce approaches that include ethical aspects when teaching and training AI (e.g., in higher education institutions, such as computer sciences).	Forsyth et al., 2021; Kowch, 2019; McDonald & Pan, 2020; Raji et al., 2021; Schneider et al., 2023*	5	1
Theorizing AI ethics	Articles in this category mainly develop new theories by taking a philosophical lens.	Allen et al., 2000; Burema et al., 2023; Seo & Thorson, 2022*; Terzis, 2020	4	1

Considering the impact of ethical theory on how researchers address the concept of moral agency, future work on AI ethics must involve different ethical theories. As Allen et al. (2000, p. 253) stated: “There is, then, a clear sense of ‘morally good’ which a utilitarian can apply to an agent’s actions irrespective of how the agent decided upon that action (consequentialism). Not so for Kant [deontology]. According to Kant, for an action to be morally good, the action must, as he puts it, be done out of respect for the categorical imperative. Here we follow one account of the categorical imperative and assume that ‘acting out of respect for it’ means simply that the agent acted as it did because it determined that the action in question was consistent with the categorical imperative. In this sense of the term, an action cannot be morally good unless the agent, in fact, reasoned in certain fairly complex ways.”

In the “top IS” articles, 62.50% focus on applied ethics, 12.50% on consequentialism, and 6.25% each on virtue ethics and deontology. Applied ethics often underlies a deontological view that is either implicit (18.75%) or unclaimed (absent, 18.75%). While 32.73% of all 61 articles in the final sample explicitly claimed the use of ethical theory, 62.50% of the “top IS” articles explicitly outlined the underlying ethical theory. Future researchers might clarify their understanding of AI ethics, especially since there can be “...tensions between practitioners’ interpretation of ethical principles in their work and ethos tensions. In this vein, we argue that understanding the tensions that can occur in practice and how they are tackled is key to studying ethics in practice. Understanding how AI practitioners perceive and apply ethical principles is necessary for practical ethics to contribute toward an empirically grounded, Responsible AI.” (Figueras et al., 2022, p. 1).

Regarding our **second research question** on AI ethics within IS research, we see that AI ethics research, design, and management are deeply interconnected with AI regulation, education, and theorization, as detailed below.

When *researching AI ethics*, researchers often apply a societal lens, delving into themes like social justice (e.g., Greene et al., 2019), fair play (e.g., Estrada, 2018), feminism (e.g., Siapka, 2022) corresponding to metaethics, or explainable AI (e.g., Vainio-Pekka et al., 2023) and AI safety (e.g., Zhang et al., 2021) corresponding to applied ethics. In doing so, researchers favor investigations that favor breadth over depth, which might be due to the difficulties in aligning different perceptions of what is right or wrong depending on different normative ethical theories that are used within different investigational settings (Seo & Thorson, 2022). The complexity of AI ethics emerges as a recurring theme. Mirbabaie et al. (2022) and Siau and Wang (2020) point out the lack of clear definitions and conceptualizations of AI ethics and the absence of a general theory to explain its ethical dimensions from a normative point of view.

Designing ethical AI revolves around the central role of ethical principles following deontological ethics. Benner et al. (2021) and Schlimbach and Khosrawi-Rad (2022) highlight the importance of these principles, while Emdad et al. (2023), Hooker and Kim (2018), Vakkuri et al. (2021), and Vanhée and Borit (2022) focus on guidelines and frameworks. The importance of experimentation and learning ethical behavior through participatory approaches (involving human users’ virtue ethics) is also emphasized, as discussed by Eicher et al. (2018), Chaput et al. (2021), and Gerdes

(2022). This complements research on nudging (i.e., guiding individuals' behavior toward a beneficial choice for themselves or society) (Mirbabaie et al., 2022).

Managing AI ethics involves addressing different perceptions of moral agency (in particular, AMA) and, thus, how to attribute responsibility (e.g., Dyrkolbotn et al., 2018; Figueras et al., 2022; Gómez de Ágreda, 2020; Rismani & Moon, 2023; Wang et al., 2020). Research on organizational practices following the organizational stakeholders' perceptions of what is right or wrong (virtue ethics), such as the work of Agbese et al. (2023) and Akbari Ghatar et al. (2023), reveals varying practices for evaluating fairness and the difficulty of reaching a shared definition of fairness in AI ethics. This perspective is further underlined by the work of Figueras et al. (2022) and Mayer et al. (2021), who analyze the perceptions of responsible AI practices in organizations. Their findings align with the challenges of a missing shared understanding of AI ethics. A few articles specifically address managing responsible AI by aligning organizational aims of AI implementation, such as increasing working efficiency (corresponding to consequentialism) with CDR norms (corresponding to deontological ethics) and workforces and leadership values (corresponding to virtue ethics) as seen in the work of Figueras et al. (2022), Rismani and Moon (2023), and Wang et al. (2020). These studies contribute to a broader understanding of how organizations manage AI ethics.

Regulating AI ethics involves addressing its societal impact and exploring potential regulatory and legal approaches. Our review highlights a growing call for training and educating AI governance topics (Mäntymäki et al., 2023; Minkkinen & Mäntymäki, 2023). In doing so, structured approaches with defined ethical principles (corresponding to deontological ethics) might guide AI technologies' ethical use and development within organizations and at a broader societal level (corresponding to humans' virtue ethics). The emphasis on governance reflects an understanding that AI's ethical considerations are not limited to design principles (deontological ethics) but require active and ongoing management and oversight to ensure responsible use (virtue ethics). Therefore, appropriate training and education are required.

When theorizing AI ethics, sector-based perspectives are vital. Burema et al. (2023) emphasize that understanding AI systems within specific sectoral cultures is crucial, given that perceptions of responsibilities are different depending on the ethical understanding of different cultures (specifically, virtue ethics). Terzis (2020) addresses the abstract nature of AI ethics discussions and the "illusion of agreement" among stakeholders due to varying interpretations of ethical behavior based on their virtue ethics.

10 Discussion

10.1 Theoretical Implications

Our review outlines two theoretical implications (corresponding to our third RQ).

We need a more fundamental discussion on how our ethical understanding impacts our perceptions of responsibilities: Mayer et al. (2021) present a framework that shows how companies can encourage their employees to implement ethical AI and

to take responsibility for specific actions, using different combinations of implementation measures and mechanisms of AI ethics policies: employee engagement, organizational embedding, practical support, and assurance processes. Mirbabaie et al. (2022) further claim that IS research lacks a general theory and fundamental research on the ethical dimensions in the complex AI context guiding our understanding of responsibilities. In addition, Vakkuri et al. (2021) go straight to the heart of the issue and call for an in-depth examination of the term ethical AI itself. The need for a definition and an examination of the terminology is also evident as some authors (e.g., Agbese et al., 2023; Akbari Ghatar et al., 2023; Corvite et al., 2023) use the term ethical AI without defining it, which might increase readers' perceptions on an AI that might be accountable resulting in responsibility gaps (Coeckelbergh, 2020). Cooper et al. (2022) also mention that future work should increasingly address the question of which actors should take responsibility in an algorithmic society and to what extent.

We need more diverse and empirical research on how different value systems impact AI ethics principles: Empirical investigations are needed to compare AI ethics principles grounded in the value systems of various industries and company types, utilizing globally diverse sample sizes (Agbese et al., 2023; Jantunen et al., 2021; Mayer et al., 2021; Seppälä et al., 2021). For instance, publicly listed companies may be subject to more pressure from stakeholders (Mayer et al., 2021) and, therefore, implement AI ethics measures earlier. Alternatively, a geographical focus can result in a lack of generalizability as value systems differ between countries and their underlying cultures, forming their norms on what might be right or wrong guiding AI adoption guidelines (Agbese et al., 2023; Seppälä et al., 2021). Cross-sectoral (Minkkinen & Mäntymäki, 2023) and comparative studies can provide insights into industries, sectors, and regional differences, taking into account highly regulated areas such as medicine or less ethically sensitive areas such as manufacturing (Mäntymäki et al., 2023). Future studies should also compare the implications of AI ethics between the private and public sectors (Figueras et al., 2022). According to Corvite et al. (2023), developing close-ended surveys might contribute to marginalized groups or vulnerable individuals (McDonald & Pan, 2020). As AI technologies and the associated ethical considerations are developing so rapidly, researchers must continuously monitor and account for this development (Burema et al., 2023), wherefore new contexts should continuously be added (Forsyth et al., 2021; Lim & Kwon, 2021) and longitudinal studies should be implemented (Seppälä et al., 2021). Schneider et al. (2023) argue that empirical research would benefit the discourse on ethical AI.

10.2 Limitations and Future Work

Our scoping review has potential limitations. First, the search strategy did not include backward and forward searches, which might have limited the comprehensiveness of the literature review by excluding relevant articles not indexed in the four databases searched. Additionally, our search string did not encompass all potential keywords related to AI, such as “Machine Learning” (ML), which may have resulted in missing some relevant studies. Lastly, we focused on specific databases for keyword search.

Due to the interdisciplinary nature of the AI ethics field (Islam & Greenwood, 2023), future work might integrate further literature, such as from the Ethical Philosophy field. Future research might extend our insights by investigating AI ethics in the context of applied, descriptive, and meta-ethics, given the clear dependence between the theoretical justification of AI ethics and the interpretation of results.

10.3 Practical Implications

Our review highlights a gap between theory and practice (Tidjon & Khomh, 2022; Vainio-Pekka et al., 2023; Weber, 2020), which is critical as practitioners are those implementing AI solutions (Vainio-Pekka et al., 2023). Following Mirbabaie et al. (2022), the knowledge gained in the theoretical works should create a scientific foundation as AI influences society and individuals. According to the authors, the influence of NGOs and the media on AI use should also be investigated in practice. Best practices can bridge the gap from theory to practice, supporting “the creation of new methods, tools, frameworks, and guidelines to implement ethical AI principles into practice” (Seppälä et al., 2021, p. 15). This transfer is crucial since Mayer et al. (2021) underline that little is known about AI ethics in practice.

One possibility for practitioners might be the orientation on the most often used and cited ethical principles for AI. Aslan et al. (2022, p. 2) stated: “[...] the most common definitions and frameworks are AI4People (Floridi et al., 2018), the OECD Principles of AI (OECD 2019), and the Ethics Guidelines for Trustworthy AI (‘High-level expert group on Artificial Intelligence’ (HLEG; AI, 2019)).” Besides the HLEG and OECD principles, the authors of our final sample also mentioned the AI Act of the European Commission (e.g., Tidjon & Khomh, 2022; Walke et al., 2023; Westerstrand, 2023), the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (e.g., Figueras et al., 2022; Lim & Kwon, 2021; Ong, 2021), and the General Data Protection Regulation (GDPR; e.g., Almeida et al., 2020; Fabiano, 2019; Siapka, 2022).

Practitioners can also integrate AI ethics into curricula since AI-related knowledge and experiences might contribute to more responsible usage of AI (Weber et al., 2023). For instance, Henderson (2019) argues that ethical discourse should be expanded and integrated into computer science curricula. Additionally, McDonald and Pan (2020) support the integration of intersectional AI thinking into students’ curricula.

Conclusion

Our scoping review on AI ethics research synthesizes IS literature to investigate what normative ethical theories are most commonly used, what topics of AI ethics are studied most and calls for future work in this field. The review demonstrates that IS research lacks the integration of several views on normative ethics, resulting in assessing ethics mainly from the AI’s perspective by integrating ethical principles into AI’s design, rather than considering how a human user’s virtue equally impacts human-AI implications (e.g., through their data input and the way they critically reflect and use AI outputs).

References

- Agbese, M., Mohanani, R., Khan, A., & Abrahamsson, P. (2023), 'Implementing AI Ethics: Making Sense of the Ethical Requirements', in 'Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering', pp. 62–71.
- Akbari Ghatari, P., Pappas, I., & Vassilakopoulou, P. (2023), 'Practices for Responsible AI: Findings from Interviews with Experts', in 'AMCIS 2023 Proceedings'.
- Allen, C., Varner, G., & Zinser, J. (2000), 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence*, **12**(3), pp. 251–261.
- Almeida, P., Santos, C., & Farias, J. S. (2020), 'Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance', in 'Proceedings of the Hawaii International Conference on System Sciences (HICSS) 2020'.
- Arksey, H., & O'Malley, L. (2005), 'Scoping studies: Towards a methodological framework', *International Journal of Social Research Methodology*, **8**(1), pp. 19–32.
- Aslan, A., Greve, M., & Lembcke, T.-B. (2022), 'Let's Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence', in 'AMCIS 2022 Proceedings'.
- Auld, G., Casovan, A., Clarke, A., & Faveri, B. (2022), 'Governing AI through ethical standards: Learning from the experiences of other private governance initiatives', *Journal of European Public Policy*, **29**(11), pp. 1822–1844.
- Baird, A., & Maruping, L. (2021), 'The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts', *Management Information Systems Quarterly*, **45**(1), pp. 315–341.
- Bankins, S., & Formosa, P. (2023), 'The Ethical Implications of Artificial Intelligence (AI) for Meaningful Work', *Journal of Business Ethics*, **4**, pp. 1–16.
- Beauchamp, T. L., & Bowie, N. E. (1988), *Ethical Theory and Business* (Subsequent Edition), Prentice Hall.
- Behdadi, D., & Munthe, C. (2020), 'A Normative Approach to Artificial Moral Agency', *Minds and Machines*, **30**(2), pp. 195–218.
- Benner, D., Schöbel, S., & Janson, A. (2021), 'It is only for your own good, or is it? Ethical Considerations for Designing Ethically Conscious Persuasive Information Systems', in 'AMCIS 2021 Proceedings'.
- Bentham, J. (2007), *An Introduction to the Principles of Morals and Legislation*, Dover Publications Inc.
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021), 'Managing Artificial Intelligence', *MIS Quarterly*, **45**, pp. 1433–1450.
- Bilal, A., Wingreen, S., Sharma, R., & Jahanbin, P. (2021), 'Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics', in 'ACIS 2021 Proceedings'.
- Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021), 'Ethical Management of Artificial Intelligence', *Sustainability*, **13**(4), Article 4.
- Brown, L., & Aristotle. (2009), *The Nicomachean Ethics* (D. Ross, Trans.; New Edition).
- Burema, D., Debowski-Weimann, N., von Janowski, A., Grabowski, J., Maftai, M., Jacobs, M., van der Smagt, P., & Benbouzid, D. (2023), 'A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context', in 'Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society', pp. 705–714.
- Champagne, M., & Tonkens, R. (2023), 'A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot harm', *Science and Engineering Ethics*, **29**(4), p. 27.
- Chaput, R., Duval, J., Boissier, O., Guillermin, M., & Hassas, S. (2021), 'A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior', in 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 13–23.
- Coeckelbergh, M. (2020), *AI Ethics*, The MIT Press.
- Colmenarejo, A. B., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., & Kinder-Kurlanda, K. (2022), 'Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation', in 'Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society', pp. 107–118.
- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021), 'Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context', *Ethics and Information Technology*, **23**, pp. 803–814.
- Cook, T. (2023), 'Robust Artificial Moral Agents and Metanormativity', in 'Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society', pp. 162–169.
- Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022), 'Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning', in '2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 864–876.
- Corvite, S., Roemmich, K., Rosenberg, T. I., & Andalibi, N. (2023), 'Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens', in 'Proceedings of the ACM on Human-Computer Interaction', **7**(CSCW1), pp. 124:1–124:38.
- Dignum, V. (2019), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer Verlag.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023), 'Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on

- opportunities, challenges and implications of generative conversational AI for research, practice and policy', *International Journal of Information Management*, **71**, 102642.
- Dyrkolbotn, S., Pedersen, T., & Slavkovik, M. (2018), 'On the Distinction between Implicit and Explicit Ethical Agency', in 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 74–80.
- Eicher, B., Polepeddi, L., & Goel, A. (2018), 'Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies', in 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 88–94.
- Eitel-Porter, R. (2021), 'Beyond the promise: Implementing ethical AI', *AI and Ethics*, **1**(1), pp. 73–80.
- Emdad, F. B., Ho, S., Ravuri, B., & Hussain, S. (2023), 'Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence', in 'AMCIS 2023 Proceedings'.
- Estrada, D. (2018), 'Value Alignment, Fair Play, and the Rights of Service Robots', in 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 102–107.
- Fabiano, N. (2019), 'Robotics, intelligent systems, ethics and data protection: The new challenge', in 'Proceedings of the 2nd International Conference on Applications of Intelligent Systems', pp. 1–5.
- Figueras, C., Verhagen, H., & Pargman, T. C. (2022), 'Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations', *Scandinavian Journal of Information Systems*, **34**(2).
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018), 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines*, **28**(4), pp. 689–707.
- Forsyth, S., Dalton, B., Foster, E. H., Walsh, B., Smilack, J., & Yeh, T. (2021), 'Imagine a More Ethical AI: Using Stories to Develop Teens' Awareness and Understanding of Artificial Intelligence and its Societal Impacts', in '2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)', pp. 1–2.
- Gaus, G. (2001), 'What is Deontology? Part One: Orthodox Views', *The Journal of Value Inquiry*, **35**, pp. 27–42.
- Gerdes, A. (2022), 'A participatory data-centric approach to AI Ethics by Design', *Applied Artificial Intelligence*, **36**(1), p. 2009222.
- Gewirth, A. (1960), 'Meta-Ethics and Normative Ethics', *Mind*, **69**(274), pp. 187–205.
- Giermndi, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzeqi, A. (2022), 'The dark sides of people analytics: Reviewing the perils for organisations and employees', *European Journal of Information Systems*, **31**(3), pp. 410–435.
- Gómez de Ágreda, Á. (2020), 'Ethics of autonomous weapons systems and its applicability to any AI systems', *Telecommunications Policy*, **44**(6), p. 101953.
- Greene, D., Hoffmann, A. L., & Stark, L. (2019), 'Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Artificial Intelligence and Machine Learning', in 'Proceedings of the Hawaii International Conference on System Sciences 2019 (HICSS)'.
- Ha, Y. J. (2022), 'South Korean Public Value Coproduction Towards 'AI for Humanity': A Synergy of Sociocultural Norms and Multistakeholder Deliberation in Bridging the Design and Implementation of National AI Ethics Guidelines', in 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 267–277.
- Hawkins, W., & Mittelstadt, B. (2023), 'The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices', in '2023 ACM Conference on Fairness, Accountability, and Transparency', pp. 261–270.
- Henderson, T. (2019), 'Teaching Data Ethics: We're going to ethics the heck out of this', in 'Proceedings of the 3rd Conference on Computing Education Practice', pp. 1–4.
- Heyder, T., Passlack, N., & Posegga, O. (2023), 'Ethical management of human-AI interaction: Theory development review', *The Journal of Strategic Information Systems*, **32**(3), p. 101772.
- HLEG; AI. (2019), 'Ethics guidelines for trustworthy AI', <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>, [last visited June 28, 2024].
- Hooker, J. N., & Kim, T. W. N. (2018), 'Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic', in 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 130–136.
- Islam, G., & Greenwood, M. (2023), 'Ethical Research in Business Ethics', *Journal of Business Ethics*, **182**(1), pp. 1–5.
- Jantunen, M., Halme, E., Vakkuri, V., Kemell, K.-K., Rousi, R., Mikkonen, T., Duc, A. N., & Abrahamsson, P. (2021), 'Building a Maturity Model for Developing Ethically Aligned AI Systems', *Selected Papers of the IRIS*, Issue Nr 12 (2021).
- Jobin, A., Ienca, M., & Vayena, E. (2019), 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, **1**(9), Article 9.
- Kant, I. (1797), *Die Metaphysik der Sitten*, Nicolovius.
- Kellogg, K. C., Valentine, M. A., & Christin, A. (2020), 'Algorithms at Work: The New Contested Terrain of Control', *Academy of Management Annals*, **14**(1), pp. 366–410.
- Khan, A. A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., & Akbar, M. A. (2021), 'Ethics of AI: A Systematic Literature Review of Principles and Challenges (arXiv:2109.07906)', *arXiv*.
- Kowch, E. G. (2019), 'Ethics to Prepare Educators for Professional Service Robots in Classrooms', in '2019 International Joint Conference on Information, Media and Engineering (IJCIME)', pp. 478–484.
- Lim, J. H., & Kwon, H. Y. (2021), 'A Study on the Modeling of Major Factors for the Principles of AI Ethics', in 'DG.O2021: The 22nd Annual International Conference on Digital Government Research', pp. 208–218.
- Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021), 'Corporate digital responsibility', *Journal of Business Research*, **122**, pp. 875–888.

- Lowry, P. B., Moody, G. D., Gaskin, J., Galletta, D. F., Humpherys, S., Barlow, J. B., & Wilson, D. W. (2013), 'Evaluating Journal Quality and the Association for Information Systems Senior Scholars' Journal Basket Via Bibliometric Measures: Do Expert Journal Assessments Add Value?', *Management Information Systems Quarterly*, **37**(4), 993–1012.
- Mäntymäki, M., Minkkinen, M., Zimmer, M., Birkstedt, T., & Viljanen, M. (2023), 'Designing an AI governance framework: From research-based premises to meta-requirements', in 'ECIS 2023 Proceedings'.
- Mayer, A.-S., Haimel, A., Strich, F., & Fiedler, M. (2021), 'How corporations encourage the implementation of AI ethics', in 'ECIS 2021 Proceedings'.
- Mayring, P. (2014), *Qualitative content analysis—Theoretical foundation, basic procedures and software solution*, Klagenfurt.
- McDonald, N., & Pan, S. (2020), 'Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact', in 'Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2)', pp. 147:1-147:19.
- McHugh, M. L. (2012), 'Interrater reliability: The kappa statistic', *Biochemia Medica*, **22**(3), pp. 276–282.
- Mihale-Wilson, C., Hinz, O., van der Aalst, W., & Weinhardt, C. (2022), 'Corporate Digital Responsibility', *Business & Information Systems Engineering*, **64**(2), pp. 127–132.
- Mill, J. S. (2009), *Utilitarismus*, Felix Meiner Verlag.
- Minkkinen, M., & Mäntymäki, M. (2023), 'The Institutional Logics Underpinning Organizational AI Governance Practices', in 'Proceedings of the 14th Scandinavian Conference on Information Systems'.
- Mirbabaie, M., Brendel, A., & Hofeditz, L. (2022), 'Ethics and AI in Information Systems Research', *Communications of the Association for Information Systems*, **50**.
- Mueller, B. (2022), 'Corporate Digital Responsibility', *Business & Information Systems Engineering: The International Journal of WIRTSCHAFTSINFORMATIK*, **64**(5), pp. 689–700.
- Ong, D. C. (2021), 'An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence (arXiv:2107.13734)', *arXiv*.
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015), 'Synthesizing information systems knowledge: A typology of literature reviews', *Information & Management*, **52**(2), pp. 183–199.
- Passlack, N., Hammerschmidt, T., Klemm, F., & Posegga, O. (2023), 'How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success', in 'ICIS 2023 Proceedings'.
- Polyviou, A., & Zamani, E. (2022), 'Towards Europe's AI Strategy: Desires & Realities', in 'ECIS 2022 Proceedings'.
- Rai, A., Constantinides, P., & Sarker, S. (2019), 'Editor's Comments: Next-Generation Digital Platforms: Toward Human-AI Hybrids', *Management Information Systems Quarterly*, **43**(1), pp. iii–ix.
- Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021), 'You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education', in 'Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency', pp. 515–525.
- Rismani, S., & Moon, A. (2023), 'What does it mean to be a responsible AI practitioner: An ontology of roles and skills', in 'Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society', pp. 584–595.
- Schlimbach, R., & Khosrawi-Rad, B. (2022), 'Towards Ethical Design Features for Pedagogical Conversational Agents', in 'AMCIS 2022 Proceedings'.
- Schneider, J., Meske, C., & Bikic, A. (2023), 'How individuals can shape AI through data—An AI literacy and morality perspective', in 'ECIS 2023 Proceedings'.
- Seo, H., & Thorson, S. (2022), 'Computation, Rule Following, and Ethics in AIs', in 'Proceedings of the Hawaii International Conference on System Sciences 2022 (HICSS)'.
- Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021), 'From Ethical AI Principles to Governed AI', in 'ICIS 2021 Proceedings'.
- Seymour, M. (2018), 'Artificial Intelligence Is No Match for Human Stupidity: Ethical Reflections on Avatars and Agents', in 'ACIS 2018 Proceedings'.
- Siapka, A. (2022), 'Towards a Feminist Metaethics of AI', in 'Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society', pp. 665–674.
- Siau, K., & Wang, W. (2020), 'Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI', *Journal of Database Management (JDM)*, **31**(2), pp. 74–87.
- Søvik, A. O. (2022), 'What overarching ethical principle should a superintelligent AI follow?', *AI & Society*, **37**(4), pp. 1505–1518.
- Stolz, K., Hammerschmidt, T., & Posegga, O. (2024), 'How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers toward Sustainable Consumption: Exploring Nudges for Green Decision-Making', in 'Proceedings of the Hawaii International Conference on System Sciences 2024 (HICSS)'.
- Sullivan, Y. W., & Wamba, S. F. (2022), 'Moral Judgments in the Age of Artificial Intelligence', *Journal of Business Ethics*, **178**(4), pp. 917–943.
- Susser, D. (2019), 'Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures', in 'Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society', p. 1.
- Templier, M., & Paré, G. (2018), 'Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals', *European Journal of Information Systems*, **27**(5), pp. 503–550.
- Terzis, P. (2020), 'Onward for the freedom of others: Marching beyond the AI ethics', in 'Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency', pp. 220–229.
- Tidjon, L. N., & Khomh, F. (2022), 'The Different Faces of AI Ethics Across the World: A Principle-Implementation Gap Analysis (arXiv:2206.03225)', *arXiv*.
- Tóth, Z., Caruana, R., Gruber, T., & Loebbecke, C. (2022), 'The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability', *Journal of Business Ethics*, **178**(4), pp. 895–916.

- Unver, M. B. (2023), 'Rebuilding "ethics" to govern AI: How to re-set the boundaries for the legal sector?', in 'Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law', pp. 306–315.
- Vainio-Pekka, H., Agbese, M. O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023), 'The Role of Explainable AI in the Research Field of AI Ethics', *ACM Transactions on Interactive Intelligent Systems*.
- Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021), 'ECCOLA—A method for implementing ethically aligned AI systems', *Journal of Systems and Software*, **182**, p. 111067.
- Vallor, S. (2016), *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press.
- Vanderelst, D., & Winfield, A. (2018), 'The Dark Side of Ethical Robots', in 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 317–322.
- Vanhée, L., & Borit, M. (2022), 'Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence', *Journal of Artificial Intelligence Research*, **73**, pp. 619–631.
- Walke, F., Bennek, L., & Winkler, T. (2023), 'AI in Government: A Study on Explainability of High-Risk AI-Systems in Law Enforcement & Police Service', in 'Wirtschaftsinformatik 2023 Proceedings'.
- Walstrom, K. A., & Hardgrave, B. C. (2001), 'Forums for information systems scholars: III', *Information & Management*, **39**(2), 117–124.
- Wang, Y., Xiong, M., & Olya, H. (2020), 'Toward an Understanding of Responsible Artificial Intelligence Practices', in 'Proceedings of the Hawaii International Conference on System Sciences 2020 (HICSS)'.
- Weaver, G. R., & Trevino, L. K. (1994), 'Normative and Empirical Business Ethics: Separation, Marriage of Convenience, or Marriage of Necessity?', *Business Ethics Quarterly*, **4**(2), pp. 129–143.
- Weber, M. (1988), *Gesammelte Aufsätze zur Wissenschaftslehre*, Mohr Siebeck GmbH & Co. KG.
- Weber, P., Pinski, M., & Baum, L. (2023), 'Toward an Objective Measurement of AI Literacy', in 'PACIS 2023 Proceedings'.
- Weber, R. H. (2020), 'Socio-ethical values and legal rules on automated platforms: The quest for a symbiotic relationship', *Computer Law & Security Review*, **36**, p. 105380.
- Westerstrand, S. (2023), 'Ethics in the intersection of AI and democracy: The AIDEM Framework', in 'ECIS 2023 Proceedings'.
- Wright, S. A. (2020), 'AI in the Law: Towards Assessing Ethical Risks', in '2020 IEEE International Conference on Big Data (Big Data)', pp. 2160–2169.
- Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018), 'Building Ethics into Artificial Intelligence (arXiv:1812.02953)', *arXiv*.
- Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021), 'Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers (arXiv:2105.02117)', *arXiv*.

Paper III

Hammerschmidt, T.; Passlack, N., Stolz, K. & Posegga, O. (2024). *AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitude Toward AI Assistants?* Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, <https://hdl.handle.net/10125/106426>.

AI Narratives: What Can They Tell Us About Individuals' AI Literacy and Emotional Attitudes toward AI Assistants?

Research Paper – Accepted Version

Teresa Hammerschmidt
University of Bamberg
teresa.heyder@uni-bamberg.de

Nina Passlack
University of Bamberg
nina.passlack@uni-bamberg.de

Katharina Stolz
University of Stuttgart
katharina.stolz@bwi.uni-stuttgart.de

Oliver Posegga
University of Bamberg
oliver.posegga@uni-bamberg.de

Abstract

How individuals understand Artificial Intelligence (AI) affects whether they can interact with AI assistants appropriately. To foster the appropriate use of AI assistants, individuals require realistic perceptions of what AI can or cannot do. However, perceptions (which we refer to as AI narratives) depend on individuals' AI literacy and their emotional attitudes regarding AI assistants. To investigate how literate individuals are and their emotional attitudes when dealing with AI assistants, we suggest developing a better understanding of their different AI narratives. Through a qualitative online survey, we explore differences in AI narratives among individuals with positive, ambivalent, or negative emotional attitudes regarding AI and among those with low, medium, or high levels of AI literacy. This work provides two research-guiding propositions on an individual's AI understanding and two recommendations for managing realistic AI perception-building.

Keywords: AI narratives, AI perceptions, AI attitudes, AI literacy, human-AI interaction

1. Introduction

Recent developments in Artificial Intelligence (AI), especially through the growing maturity of Large Language Models (LLMs, such as GPT4), offer tremendous potential for individuals to become more efficient in daily tasks and job assignments (Benbya et al., 2021). Hence, the use of generative AI assistants for personal and work purposes is increasing (Dwivedi et al., 2023). With their ability “to perform cognitive functions that we associate with human minds” (Rai et al., 2019, p. iii), AI assistants possess greater autonomy and interactivity compared to earlier forms of digital assistants (Dwivedi et al., 2023; Maedche et al., 2019). These newer AI assistants, such as voice- (e.g., Alexa or Siri) and text-based assistants (e.g., ChatGPT or DALL-E), transform how individuals perceive and interact with them (Diederich et al., 2022; Dwivedi et al., 2023; Maedche et al., 2019).

Many individuals use AI assistants, but most do not adequately understand what it is with which they are interacting (Baidoo-Anu & Owusu Ansah, 2023; Dwivedi et al., 2023). The human-like behavior of AI assistants and their self-learning abilities (Dwivedi et al., 2023; Glikson & Woolley, 2020; Greene et al., forthcoming) can lead to misunderstanding, unrealistic perceptions, and misconceptions about what AI is, what it can do, and how it can be used (Diederich et al., 2022; Jordan, 2019; Willcocks, 2020). Individuals must develop realistic perceptions of AI (Diederich et al., 2022) to ensure that they interpret AI assistants' abilities correctly and, thus, use AI appropriately, that is, in a mindful and responsible manner that includes considering the possible negative consequences for society and the quality of life (Carolus et al., 2023; Kuzior & Kwilinski, 2022).

To develop realistic perceptions of AI, it is important to improve individual knowledge of AI assistants (Diederich et al., 2022). Individual's AI knowledge and the ability to understand and use AI correctly can be described with the concept of AI literacy, that is, the “new competencies [that] will be necessary in a future in which AI transforms the way that we communicate, work, and live with each other” (Long & Magerko, 2020, p. 598).

Beyond their literacy, individuals' emotional attitudes can lead to utopian or dystopian AI perceptions of AI and thus prevent them from using AI assistants appropriately (Cave & Dihal, 2019). Media and public discourse, with hopeful and frightening stories used to evoke the emotional attitudes of individuals, influence how individuals develop perceptions of AI (Cave & Dihal, 2019; Chubb et al., 2022). Research itself often “seem[s] to polarize around two storylines – hype or fear” (Willcocks, 2020, p. 287). This can result in individuals perceiving that “AI [will] become more powerful than humans” (Cave et al., 2019, p. 331). In short, both public and academic discourse can affect individual AI perceptions by triggering negative or positive emotional attitudes toward AI (Chubb et al., 2022).

To avoid unrealistic AI perceptions, research needs to understand how literate and emotional individuals perceive AI assistants (Diederich et al., 2022; Dwivedi et al., 2023). Individuals use narratives to express their perceptions of AI (Cave & Dihal, 2019), that is, stories through which they communicate and visualize their

thoughts, feelings, experiences, and reflections (Schiff, 2012). Individual literacy, *what people know and think*, and individual emotional attitudes, *what people feel*, affect each other (Pinski et al., 2023). We expect both to influence AI narratives.

Research to date has focused mainly on dichotomous and collective narratives of “good” and “bad” rather than taking a more nuanced view of the “gap between dominant narratives” from an individual’s perspective (Chubb et al., 2022, p. 1). We aim to explore differences in the narratives of individuals at various levels of literacy with AI assistants and with various emotional attitudes, following the call of Koukouvouinou and Holmström (2022, p. 13) to further investigate differences between utopic and dystopic narratives. Our research question is: *How do AI narratives differ among individuals depending on their AI literacy and emotional attitudes regarding AI assistants?*

To answer this question, we conducted an exploratory qualitative study based on an online survey with open-ended questions. Exploratory research is appropriate, as it uses data for the “reconciliation of polyphonic narrative[s]” mainly based on inductive reasoning and explores insights (Sarker et al., 2018, pp. 764–765). We asked participants about their emotional attitudes toward AI assistants and their AI literacy to cluster them into similar groups. Within those groups, we explored their narratives inductively by asking open-ended questions about their views of AI and its impact. Following Sarker et al. (2018, p. 762), we used theoretical background on AI narratives, literacy, and emotional attitudes “up-front to guide the design and execution” of our exploratory study. Thereby, we aim to provide a more nuanced analysis of AI narratives to increase understanding of how and why AI perceptions might differ depending on individuals’ emotional attitudes and AI literacy. We hope this will contribute to the understanding of what is required to foster individual development of more realistic perceptions of AI and thus help people interact with AI appropriately.

Next, we present the theoretical background, followed by the method. The results are then followed by the discussion. The paper ends with a conclusion.

2. Theoretical background

This section provides our theoretical background on AI narratives, literacy, and emotional attitudes.

2.1. AI narratives

Narratives are innumerable, as, by definition, all human forms of expression are narratives or can be treated as such (Czarniawska, 2009). Research often distinguishes between strictly negative and strictly positive narratives. The conditions of individuals’ context, such as culture, media coverage, and education, may create a utopia in one context and a dystopia in another (Cave et al., 2019). AI narratives are highly polarized (Cave et al., 2019); with respect to AI, positive expressions often refer to hopes or utopias and negative expressions to fears or dystopias (Cave et al., 2019; Fast & Horvitz, 2017).

Table 1 introduces typical AI narratives that have been discussed in research to date; it follows the nine perception categories of Diederich et al. (2022, p. 100). For each category, we used an example from our initial literature screening within the AIS Electronic Library and Google Scholar, in which we searched for the following keywords: (“Artificial Intelligence”) AND (“narrative*” OR “perception*”). We do not claim comprehensiveness in our work (Sarker et al., 2018); rather, we use “some theory ... up-front to guide the design and execution” of our exploratory study (Sarker et al., 2018, p. 762).

Table 1. Categories of existing AI narratives.

Category	Hopes	Fears
Perception (e.g., humanness)	“Humans merge with AI in a positive way, e.g., robotic limbs for the disabled, positive discussions about potential rise of transhumanism” (Fast & Horvitz, 2017, p. 964)	“Humans merge with AI in a negative way, e.g., cyborg soldiers.” (Fast & Horvitz, 2017, p. 964)
Acceptance (e.g., usefulness)	“AI might make our day-to-day lives easier because we could ask computers to do more tasks for us.” (Cave et al., 2019, p. 333)	“... an enlargement often results in [AI] solutions that are brittle and that become useless as soon as the application requirements change only slightly.” (Weiss, 1999, p. 7)
Attitude (e.g., satisfaction)	“AI makes human work easier or frees us from needing to work at all, e.g., by managing our schedules, automating chores via robots” (Fast & Horvitz, 2017, p. 964)	“AI displaces human jobs, e.g., large-scale loss of jobs by blue collar workers.” (Fast & Horvitz, 2017, p. 964)
Performance (e.g., productivity)	“AI improves how [people] learn, e.g., through automatic tutoring or grading, or providing other kinds of personalized analytics.” (Fast & Horvitz, 2017, p. 964)	“AI kills people or leads to instabilities and warfare through military applications, e.g., robotic soldiers, killer drones.” (Fast & Horvitz, 2017, p. 964)
Emotion (e.g., humor)	“AI might become the perfect friend, there to listen whenever we need and ready to meet our every desire.” (Cave et al., 2019, p. 333)	“AI might cater to all our desires so well that we prefer AI interaction to human interaction.” (Cave et al., 2019, p. 333)

Trust (e.g., risk, security)	“AI might help strengthen our military power because it could provide smarter weapons.” (Cave et al., 2019, p. 333)	“AI might enable computers to become more powerful than us.” (Cave et al., 2019, p. 333)
Learning (e.g., progress)	“From reaching human ability, it will move on to superintelligence in 2 years (10%) to 30 years (75%)” (Müller & Bostrom, 2016, p. 14)	“AI systems will probably (over 50%) reach overall human ability by 2040-50, and very likely (with 90% probability) by 2075.” (Müller & Bostrom, 2016, p. 14)
Ethics (e.g., ethical behavior)	“AI enhances the health and well-being of people.” (Fast & Horvitz, 2017, p. 964)	“AI lacks ethical reasoning, leading to negative outcomes, e.g., loss of human life.” (Fast & Horvitz, 2017, p. 964)
Relationship (e.g., trust, responsibility)	“AI or expert systems help us make better decisions, e.g., when to take a meeting, or case-based reasoning for business executives” (Fast & Horvitz, 2017, p. 964)	“Humans lose control of powerful AI systems” (Fast & Horvitz, 2017, p. 964)

This categorization of positive and negative expressions of narratives requires a more nuanced analysis given that AI perceptions of individuals in reality occur in mixed forms and can vary as a consequence of human-AI collaboration (Chubb et al., 2022). Hence, we aim to give nuance to these bipolar and dominant AI narratives by exploring possible differences between individuals’ narratives depending on their different levels of AI literacy and emotional attitudes. We introduce this in the following, using the narrative categories in Table 1 to structure participants’ responses deductively.

As argued in the introduction, previous research demonstrates that AI narratives are interrelated with AI literacy and emotional attitudes. “It seems that the perceived level of machine intelligence moderates not only the steepness of the trust trajectory but also the activities and psychological perceptions that lead to cognitive trust” (Glikson & Woolley, 2020, p. 638). While how literate individuals are can help in developing realistic perceptions in terms of *what people know and think*, and thus can enhance responsible and mindful use of AI through AI knowledge (Carolus et al., 2023; Diederich et al., 2022), individual emotional attitudes – that is, *what people feel* – can lead to perceptions overly utopian or dystopian perceptions of AI, preventing those individuals from using AI assistants appropriately (Cave & Dihal, 2019). This illustrates that AI literacy and emotional attitudes are highly interrelated (Pinski et al., 2023). Fostering AI literacy can help balance individuals’ emotions and attitudes toward AI, which can motivate them to acquire further capabilities for using AI appropriately (Ng et al., 2021).

2.2 AI literacy

Beyond its use in the education context and its meaning with respect to the specific ability to read, literacy as we use it here comprises capabilities, such as knowledge, competencies, and skills, that are required for interaction and participation with other humans or artifacts (Barton, 2001). The concept of literacy has changed in light of technological progress (McLean, 2013), and “digital literacy” is often used to describe the capabilities required in individuals to decrease inequalities in the use of technology (Tapashi, 2018); this corresponds to the idea of literacy empowering people to create social equality. Gilster (1997, p. 1) introduced digital literacy as “the ability to understand and use information in multiple formats from a wide range of sources when it is presented via computers,” after which many definitions followed.

AI-based technologies are “no longer always subordinate to a human agent, [as they] can now assume responsibility for tasks” (Baird & Maruping, 2021, p. 315) by acting with greater independence (Dwivedi et al., 2023). Current theories on literacy have become inadequate to account for this transformation in human-machine interaction and the relationships it creates (Leander & Burriss, 2020). In response, Long and Magerko (2020) introduced AI literacy as the competencies required to interact with AI successfully, and some conceptualizations of AI literacy followed in Information Systems (IS) research (e.g., Cetindamar et al., 2022; Ng et al., 2021).

While all these conceptualizations structure the relevant aspects of AI literacy in different ways, they have in common an emphasis on the increasing relevance of ethical abilities (e.g., judgment, critical thinking) that result from the greater autonomy of AI-based technologies and how this autonomy affects human-AI interaction (Baird & Maruping, 2021). Ethical abilities enable individuals to take account of ethical considerations with respect to AI, helping them better assess their own ethical perceptions, which are related to subjective moral values and norms, in the context of human-AI interaction (Ng et al., 2021). Hence, for realistic perceptions of AI and mindful use of AI (Carolus et al., 2023), individuals must also be aware of ethical considerations (Diederich et al., 2022). By conducting a systematic literature review and expert interviews, Pinski and Benlian (2023) conceptualized AI literacy and proposed a measurement instrument involving such ethical considerations. We used their measurement instrument to investigate the AI literacy of our participants, structuring them into groups based on possessing a high, medium, or low level of AI literacy so we could undertake a more detailed analysis of differences in their AI narratives.

2.3 Emotional attitudes toward AI

Attitudes are “a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” (Eagly & Chaiken, 2007, p. 1). Research mainly discusses three attitude components for evaluation: the cognitive, affective, and behavioral components. Emotional attitudes, which are part of the affective component (Maier et al., 2019), can be characterized as beliefs or feelings that occur in a “mental state – conscious[ly] or unconscious[ly]” as settled opinions developed through cumulative experiences of individuals whose behavior is then influenced by those experiences (Altmann, 2008, p. 146). Experiences with AI can evoke emotional attitudes, which are often ambivalent, and can involve both “strong positive and negative evaluations at the same time” (Maier et al., 2019, p. 1). For instance, overestimation can lead to a perception of unfairness and thus evoke rather negative emotional attitudes (Hsu et al., 2021). We suggest that critical reflection concerning the pros and cons of AI and its output, which can result in different emotional attitudes, is necessary to develop realistic perceptions of AI assistants, given that trust can affect behavior and individuals’ understanding of how AI acts compared to earlier technologies (Glikson & Woolley, 2020). We used the measurement instrument of Maier et al. (2019), which involves attitudinal ambivalence, and adapted it to our AI assistant context. We investigated participants’ emotional attitudes to structure them into groups with positive, ambivalent, or negative attitudes for a more nuanced analysis of their narratives regarding AI assistants.

3. Method

This section provides methodological insights into our explorative study using a qualitative online survey. As stated in the introduction, exploratory studies can be used to inductively unify and structure narratives of qualitative data, which we acquired through a qualitative online questionnaire (Sarker et al., 2018). We used the concepts introduced in the theoretical background section to guide the development of survey questions and our data analysis. Below, we explain the sample, the questionnaire design, and the data analysis.

3.1 Sample

We chose several online channels to acquire participants from a variety of backgrounds to increase sampling quality (Brace, 2018, p. 1): via social media (35.62%); e-mail (18.72%); and crowdsourcing platforms such as Clickworkers (45.66%). This took place in February 2023. Given that our heterogeneous sample was acquired randomly through diverse channels, we adopted a convenient way to label our weakly defined group of survey participants (Leiner, 2014). We clarified to potential participants that the survey was about AI assistants and provided examples. Some 219 participants completed the questionnaire, of whom we used 151. We deleted data of participants who gave incomplete or invalid answers in the free text fields ($n=12$; e.g., by entering “...” or “no answer”). Further, we considered data only from participants with experience with AI assistants ($n=25$ were eliminated) and further confirmed their experience with at least medium- to high-level AI awareness ($n=31$ were eliminated).

The final sample comprises 58.28% male and 37.09% female participants (4.64% no answer) in the following age groups: 20–29, 27.81%; 30–39, 33.11%; 40 and older, 34.44% (4.64% did not answer). The highest level of education attained are as follows: high school degree, 22.52%; bachelor’s degree, 16.56%; master’s degree, 31.79%; and 29.41% have other degrees, such as a PhD, or vocational training. Most of our participants, 66.89%, are professional workers; 10.60% are students; and 10.60% are freelancers; the remainder are retirees or did not provide an answer). Participants work in the information technology (IT) industry (13.25%); 8.61% in healthcare; 6.62% in education; 5.96% each in both finance and media; 5.30% in logistics; and 54.30 % in other industries (e.g., tourism, construction). Some 92.72% of participants completed the questionnaire in German and the remainder (7.28%) did so in English.

3.2 Questionnaire design

The questionnaire contains three sections. It begins by asking for demographic information; the first section also includes our first aptitude check, in the form of questions regarding the participants’ experience with using AI assistants. We also asked participants to provide an example of the AI assistants they typically use. The second section comprises nine qualitative, open-ended questions (see Table 2), which we adopted from Lai et al. (2020) for the explorative analysis of AI narratives. Participants had already been told that the study concerned AI assistants, so we asked about AI in general in this section, rather than about AI assistants specifically. We did this to avoid participant bias and prompting them to think more broadly about “intelligence”; we wanted them to think about concrete assistants with which they had interacted without restricting their answers to those assistants.

Table 2. Qualitative questions.

No	Question

1	How do you understand AI?
2	How would you describe the unique features of AI?
3	How do you understand Human Intelligence (HI) in comparison?
4	How would you describe the unique features of HI?
5	What associations do you make when thinking about AI?
6	Can you explain how your profession is or might be changing through AI?
7	What would be necessary to make you able to deal with AI successfully?
8	Is there a need to set limits on AI? Why or why not?
9	What impact can humans have on AI-based changes?

The third section comprises the measures for AI literacy (from Pinski and Benlian, 2023) and emotional attitudes (from the medical context-specific items for strongly negative, mixed, and strongly positive attitudes of Maier et al. (2019, p. 9), but reframed to a more general context). We used these empirically validated measures to cluster our survey participants into groups of individuals with similar literacy levels and emotional attitudes. We also gauged our participants' AI awareness based on the measures for awareness of algorithm use in Dogruel et al. (2021). This was our second aptitude check, which eliminated 31 participants who did not exhibit at least medium-level AI awareness (see sample). To reduce the impact of noise in online surveys (Treiblmaier, 2011), all quantitative constructs were measured on a five-point Likert scale ranging from "1" (strongly disagree) to "5" (strongly agree). We further provided an "I don't know" option to avoid participants' guessing. We used Cronbach's Alpha (2001) to test reliability; all measures reached the critical level of 0.7 for internal constancy (Cho & Kim, 2015). This resulted in six groups of our participants (n=151) with similar AI literacy or emotional attitudes (see Table 3).

Table 3. Groups of individuals.

	Emotional Attitudes	AI Literacy
Positive/High	n=23 (15.23%)	n=59 (39.07%)
Ambivalent/Medium	n=118 (78.15%)	n=31 (20.53%)
Negative/Low	n=10 (6.62%)	n=61 (40.40%)

The questionnaire was provided in German and English. As all measures were in English, a group of bilingual personnel translated the German version to confirm correct grammar, spelling, and content comparability between English and German answers. Before issuing the questionnaire, we conducted a pilot study (n=11) to check understandability and clarify questions (Choy, 2014).

3.3 Data analysis

Given our exploratory research design, we followed inductive coding (Mayring, 2014, pp. 80–87). Table 4 is a snapshot of inductive coding.

Table 4. Snapshot of inductive coding.

Quotes	Code	Category
"Imposing mandatory rules on AI would help prevent technology infringing human rights." I126	AI infringes on human rights	In-humanity
"I hope the AI looks at the human and does not become a tyrant" I36	AI should not endanger humans	
"AI cannot evaluate everything correctly because some things are also a matter of feeling." I41	AI has no feelings	Lack of empathy
"Although AI systems can be very powerful, they still have limitations regarding their ability to understand human emotions, creativity, and empathy." I130	AI struggles to understand human emotions	
"Especially for tasks that cannot be automated and are based on human values and emotions, the use [of AI] should be limited." I78	AI should not conduct tasks that are based on human values	Lack of morality
"I have the feeling that decisions of AIs would perhaps correspond less to my values because they always rather follow a certain logic, but this is perhaps not always correct." I11	AI may not always act according to our values	

We deductively derived AI narrative categories from our theoretical background (see Table 1) to identify patterns regarding expressions of "hope" and "fear" (Mayring, 2014, pp. 95–98). All researchers coded the

English-translated answers in Microsoft Excel. The coding quality was monitored throughout the process. We had several discussion rounds among researchers to scrutinize ideas on initial codes and to consider the results until a consensus was reached.

4. Results

This section presents the main findings of how AI narratives differ among individuals depending on their AI literacy level and emotional attitudes.

4.1 AI assistants in use

Most of our participants stated that they typically interact with text-based generative AI (88.74%); most mentioned ChatGPT. Participants using generative AI mainly demonstrated positive (15.67%) or ambivalent (77.61%) emotional attitudes. Their answers could be related to the narrative categories *acceptance* (19.84%) and *performance* (20.12%), based on the impression that AI will be a “relief from unloved administrative tasks” (I16) and is “making work easier, more efficient, and faster” (I19). They had either high (39.55%) or low (40.30%) levels of AI literacy. The remaining participants (11.26%) identified translation systems, social media algorithms, and voice agents such as Alexa or Siri as the AI-based technologies with which they typically interact. They rated themselves relatively low in AI literacy (41.18%), with ambivalent emotional attitudes (82.35%). This ambivalence can be seen in statements of participants’ associations with using AI at work. One stated, “AI offers great possibilities for learning and progress, but I think as a result, many jobs will probably be lost” (I48) and can be related mainly to the narrative categories *trust* (22.50%) and *learning* (20.00%).

4.2 Sociodemographic differences

Younger participants (20–29; 26.49%) and participants older than 50 years (18.54%) showed a higher tendency toward positive emotional attitudes (n=23) compared to those ages 30 to 50, who demonstrated somewhat ambivalent emotional attitudes (40.40%, n=118). In most cases, the answers of younger participants can be related to the narrative categories *acceptance* (26.92%) and *performance* (24.04%). In contrast, the answers of older people demonstrated a broad diversification between all narrative categories. Female participants demonstrated positive emotional attitudes (25.00%), whereas male participants had more ambivalent emotional attitudes (82.95%). At the same time, predominantly female (46.43%) and older (53.57%) participants rated their level of AI literacy as low, while male participants’ AI literacy scores were equally spread between low, medium, and high. Participants between the ages of 20 and 39 rated themselves as highly literate (52.38%). There are no significant differences with respect to participants’ educational backgrounds. However, participants working in the IT industry tended toward negative emotional attitudes (11.26%) and rated their AI literacy level as either high (45.00%) or low (35.00%). In contrast, participants in other industries such as healthcare and logistics stated that they have high levels of AI literacy (47.17%) with a greater tendency toward positive emotional attitudes (18.87%).

Table 5 is an overview of the allocation of AI literacy and emotional attitudes within the narrative categories in Table 1. Participants with low AI literacy demonstrated relatively positive (22.95%) or ambivalent (73.77%) emotional attitudes. Participants with high AI literacy mainly showed ambivalent (83.05%) emotional attitudes (equal distribution between positive at 8.48% and negative at 8.47%).

Table 5. Distribution of AI literacy and emotional attitudes within narrative categories

	n	AI literacy	Emot. Attitudes	Expression
Per-ception	38	High 39.47%	Positive 18.42%	Hope 18.42%
		Med. 18.42%	Ambival. 68.42%	Neutr. 57.89%
		Low 42.11%	Negative 13.16%	Fear 23.68%
Accep-tance	73	High 52.05%	Positive 17.81%	Hope 97.26%
		Med. 15.07%	Ambival. 76.71%	Neutr. 02.74%
		Low 32.88%	Negative 05.48%	Fear 00.00%
Attitude	9	High 66.67%	Positive 44.44%	Hope 66.67%
		Med. 11.11%	Ambival. 55.56%	Neutr. 22.22%
		Low 22.22%	Negative 00.00%	Fear 01.89%
Per-formance	69	High 43.48%	Positive 17.39%	Hope 79.71%
		Med. 20.29%	Ambival. 81.16%	Neutr. 15.94%
		Low 36.23%	Negative 01.45%	Fear 04.35%
Emotion	14	High 14.29%	Positive 42.86%	Hope 00.00%
		Med. 28.57%	Ambival. 57.14%	Neutr. 14.29%
		Low 57.14%	Negative 00.00%	Fear 85.71%
Trust	30	High 13.33%	Positive 23.33%	Hope 26.67%
		Med. 13.33%	Ambival. 70.00%	Neutr. 13.33%
		Low 73.33%	Negative 06.67%	Fear 60.00%
Learning	50	High 36.00%	Positive 08.00%	Hope 50.00%

		Med.	24.00%	Ambival.	88.00%	Neutr.	50.00%
		Low	40.00%	Negative	04.00%	Fear	00.00%
		High	00.00%	Positive	00.00%	Hope	00.00%
Ethics	2	Med.	00.00%	Ambival.	100.0%	Neutr.	50.00%
		Low	100.0%	Negative	00.00%	Fear	50.00%
		High	43.18%	Positive	09.09%	Hope	06.82%
Relation-ship	44	Med.	15.91%	Ambival.	86.36%	Neutr.	72.73%
		Low	40.91%	Negative	04.55%	Fear	20.45%

4.3 Usage of narrative categories

Table 5 shows *acceptance* and *performance* are the most used narratives, with participants who are highly AI literate. Participants had neutral perceptions regarding AI within both categories, emphasizing that “we need to weigh whether the advantages of AI outweigh the disadvantages ... we need to think about technology not only economically but also ethically” (I15). Participants with high AI literacy described the increasing use of AI at the workplace as a natural progression, and often mentioned that it required skill changes (33.15%), stating, for instance, “I think of a change in the world of work similar to what electricity brought in the last century. There will be new professions; some will change, others will remain. I see this as a normal evolutionary development” (I149). Hence, AI literate participants demonstrated realistic perceptions of AI assistants with AI narratives that balance AI’s challenges and opportunities.

In comparison, answers related to the narrative categories *emotions*, *ethics*, and *trust* came from less- literate participants, who expressed fears by expressing views that AI “replaces the human workforce” (I14), “destroys humanity [like in the] movie ‘Terminator’” (I12), represents “robots without compassion” (I134) or will lead to “dehumanization” (I69). Participants with a lower level of literacy felt AI to be dangerous for humanity because of its lack of sociability (impersonal, 5.43%) and saw AI at the workplace as replacing humans (19.02%). Moral and social deskilling concerns were raised, such as “I’m afraid we forget ... how to thank!” (I36). These participants reported that they would need education to deal with AI (37.04%), as “for many people, AI is something supernatural” (I10).

While individuals with high AI literacy reported needing “lots of practice with AI systems to get the most out of them” (I151), fewer than 10% of participants with lower AI literacy reported needing practical training to improve their AI skills, instead insisting on “AI that works intuitively” (I121) and is “user friendly” (I111). Hence, participants with low levels of AI literacy used more negative expressions in their AI narratives, insisting that an artifact’s design itself should improve human-AI interaction. It is notable that participants with lower AI literacy demonstrated more critical AI narratives (e.g., by criticizing the lack of morality of AI) than highly AI-literate participants.

4.4 Understanding of artificial and human intelligence

Participants with high AI literacy attributed cognitive intelligence mostly to AI (50.00%), describing human intelligence (HI) as multi-intelligent (44.44%). They described HI by using different facets of intelligence, including “empathy” (I7; emotional intelligence, 15.48%), “getting in touch with other humans to create community spirit” (I7; social intelligence, 9.52%), “moral decision-making” (I3; ethical intelligence, 4.76%), and “musically” (I26, creative intelligence, 9.52%). Participants with high AI literacy named features linked to intrapersonal (ethical, social, and emotional) intelligence as unique characteristics of HI.

In comparison, participants with low AI literacy described AI as more human-like (47.37%). For instance, one participant said AI is a system that “is supposed to imitate human intelligence or even better” (I100), and another described AI as “infallible, perfect, future-oriented” (I30). They expressed what were largely unrealistic expectations of how AI might change their profession (13.59%) or perceive that AI supports humans (28.80%). They did not restrict AI as cognitive intelligence and named few, if any, differences between AI and HI. Consequently, participants who are less AI literate seemed to lack a differentiated understanding of AI and HI. Even though they may employ more critical AI narratives, they seem to have more unrealistic perceptions of AI.

Looking at the understanding of AI versus HI, the findings suggest that participants with negative emotional attitudes defined HI based on the idea of multi-intelligences, using primarily ethical and emotional intelligence (55.55%). For instance, “people can make moral and emotional decisions” (I93) or “it is not only logical but also emotional and experience-based” (I20). Hence, participants with negative emotional attitudes may make more appropriate differentiations between AI and HI, while participants with positive emotional attitudes tend not to distinguish between the two in the same way.

Participants with positive emotional attitudes stated that AI could increase productivity or quality (8.33%) by attending to routine tasks (41.67%) and supporting humans (50.00%) – corresponding to the *acceptance* and *performance* narrative categories. Their associations regarding how AI affects their work were linked mainly to the fact that AI can support humans (28.80%). In comparison, those with negative attitudes often characterized AI as a trend (18.33%), such as a “hype cycle, as the value in the long-term is not clear” (I139), representing the

attitude narrative. To use AI successfully, participants with negative attitudes (33.33%) and positive attitudes (45.83%) both wanted training and education to obtain, as one put it, “a better understanding of the function and limitations of AI” (I300).

4.5 Perceived human abilities to influence AI

The relevance of ambivalent emotional attitudes can also be seen in participants’ answers regarding human influence on AI, given that all participants supported setting limits to AI’s autonomy. Participants with positive emotional attitudes would prefer setting limits (79.17%), as they expressed fear of the adverse effects of unlimited AI capabilities. For example, one stated, “Yes, I believe that AI should not take over the important things of humanity, such as form part of weapons” (I57). Those with negative emotional attitudes also asked for limits on AI (66.67%) to prevent “military misuse” (I125) and “discrimination” (I63). Only one participant with negative emotional attitudes argued against limiting AI, stating, “Setting limits to science has historically never led to the improvement of society, so no limits should be set to AI” (I150). Individuals with ambivalent emotional attitudes wanted limits to protect humans (21.74%) and avoid a loss of human capital (19.57%), stating that AI is unethical (16.95%) and imperfect (8.70%), as well as limits that would maintain humans in control of and with responsibility for AI (25.00%). Hence, participants with ambivalent emotional attitudes demonstrated greater awareness of ethical issues arising from human-AI interaction – resulting in more critical AI narratives.

Irrespective of AI literacy level, participants agreed that humans should influence AI mainly through data input (21.93%); restricted usage (27.19%); discussing and adjusting parameters for AI models (27.19%); and through regulation (15.79%). No patterns emerged regarding the emotional attitudes of participants. However, participants with a relatively low level of AI literacy reported that humans have no chance of impacting AI (6.90%).

5. Discussion

In this section, we provide theoretical implications in the form of two future research-guiding propositions and two practical implications.

5.1 Theoretical implications: Realistic AI perceptions require AI literacy

Most of our participants demonstrated ambivalent emotional attitudes, which impact the ethical perceptions of individuals interacting with AI assistants (referring to the ethical narrative category). This aligns with Maier et al. (2019) and Glikson and Woolley (2020), who identified that ambivalent emotional attitudes and emotional trust strengthen the balance between the controversy over AI’s and AI’s challenges and, thus, critical reflection on AI’s impact. The high number of our participants with ambivalent emotional attitudes, coupled with the interrelationship between emotional attitudes and AI literacy described in our theoretical background section, suggest that AI literacy might be more adequate to explain differences between individuals who develop realistic AI perceptions and those who do not. Our results showed that people with high AI literacy demonstrated more realistic perceptions of AI by weighing pros and cons, whereas people with low literacy had more unrealistic narratives. Following Diederich et al. (2022), realistic perceptions also help bring ethical considerations to human-AI interaction. Thus, we propose:

Proposition 1: *The higher the level of AI literacy, the higher an individual’s ability to develop realistic AI perceptions.*

Participants with a low level of AI literacy seemed to have issues when it came to differentiating appropriately between AI and HI. Often, participants with low literacy could not distinguish unique features of HI and tended to attribute interpersonal intelligence (e.g., social and emotional intelligence) to AI. This can be linked to the increased autonomy and human-like behavior of recent AI developments (Dwivedi et al., 2023). AI-literate individuals have more realistic narratives. Hence:

Proposition 2: *The more AI literate individuals are, the more realistic their narratives when differentiating between AI and HI.*

5.2 Practical implication: Responsible AI usage requires training

Our results highlight that AI-literate participants stated that they would require training to use AI appropriately; they saw the potential to improve responsible human-AI interaction through learning-by-doing (self-effort). Individuals with low AI literacy, in requesting greater usability, saw the potential for improving human-AI interaction within the AI artifact itself. This shows that management of the human interacting with the AI requires training for appropriate usage, just as management of the AI artifact requires developing appropriate design and implementation guidelines. This aligns with Heyder et al. (2023), who argue that thinking about ethical management of human-AI interaction requires considering both the AI and the human. “Ethical issues are particularly important [as] the inappropriate use of modern technologies can have certain negative consequences

for society. Therefore, there is a need to disseminate models of proper responsible behavior and the use of these technologies for the benefit of people and increase their quality of life” (Kuzior & Kwilinski, 2022, p. 113). Therefore, our *first recommendation*, linked to AI literacy, is to align the management of both parts of human-AI interaction – the AI and the human – to ensure responsible use of AI. Irrespective of individuals’ level of AI literacy or their emotional attitudes, education is important to understand AI’s limits. This aligns with Ng et al. (2021, p. 9), who wrote that inclusive learning “bring[s] up future responsible citizens who are component in using AI in a reliable, trustworthy and fair manner, broadening participation in AI for everyone and ensuring inclusive AI learning.”

In addition, there is a need for appropriate infrastructure and onboarding for AI assistants to foster inclusive participation. This points to one of the significant challenges of “how to design and develop inclusive AI for all” (Dwivedi et al., 2023, p. 9). Hence, to avoid a gap between individuals who can and cannot use AI appropriately, our *second recommendation*, linked to emotional attitudes, is to consider the context-specific requirements required for an inclusive human-AI interaction.

6. Conclusion

This study aimed to explore the differences in the narratives that individuals develop with respect to AI assistants depending on how literate and emotional they are, in order to provide a more nuanced analysis of AI narratives following the call of Chubb et al. (2022). We hope to increase understanding of what is required to foster realistic AI perceptions and help individuals interact with AI appropriately.

Corresponding to our research question, we determined that individuals with more positive emotional attitudes regarding AI assistants used positive expressions of narratives (e.g., hopes). Individuals with more negative emotional attitudes promoted more precise differentiations in narratives when comparing AI and HI. The study results further highlight that having an ambivalent attitude inhibits higher ethical awareness of possible issues when interacting with AI assistants that may contribute positively to the responsible usage of AI; this stands in contrast to those individuals who use AI efficiently for concrete tasks but do not consider possible biases and the moral consequences of usage. This supports the argument of Maier et al. (2019) for considering both the pros and cons of AI as part of acquiring the emotional attitudes required perceive AI realistically. Our results also demonstrate that highly AI-literate individuals have more realistic AI narratives. They described AI mainly as cognitive intelligence while using multi-intelligence to describe HI, with emotional, ethical, and social intelligence as unique features of HI. In comparison, less-AI-literate individuals did not exhibit many differences in their narratives when describing AI and HI, even though their AI narratives were rather critical.

The explorative design comes with some limitations. *First*, examining our propositions empirically by conducting a quantitative study would be essential to strengthen validity, especially with respect to possible interdependencies between AI literacy and emotional attitudes as assumed independent variables that might somehow be correlated. *Second*, our sampling could be enriched by valuable information regarding possible moderating effects of specific sociodemographic data, such as the ethical ideology of individuals, as we recognized recurring patterns in the qualitative data.

We hope our work provides valuable guidance for future research and practitioners on increasing understanding of the link between individuals’ narratives and their AI literacy and emotional attitudes regarding AI assistants.

7. References

- Altmann, T. K. (2008). Attitude: A concept analysis. *Nursing Forum*, 43(3), 144–150.
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484*.
- Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 315–341.
- Barton, D. (2001). Directions for literacy research: Analysing language and social practices in a textually mediated world. *Language and Education*, 15(2&3), 92–104.
- Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Current State and Future Opportunities. *Journal of the Association for Information Systems*, 22(2).
- Brace, I. (2018). *Questionnaire design. How to plan, structure and write survey material for effective market research*. Kogan Page Publisher.
- Carolus, A., Augustin, Y., Markus, A., & Wienrich, C. (2023). Digital interaction literacy model – Conceptualizing competencies for literate interactions with voice-based AI systems. *Computers and Education: Artificial Intelligence*, 4.
- Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary Robots". In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2), 74–78.
- Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., & Knight, S. (2022). Explicating AI literacy of employees at digital workplaces. *IEEE Transactions on Engineering Management*, 1–14.

- Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha. *Organizational Research Methods*, 18(2), 207–230.
- Choy, L. T. (2014). The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches. *IOSR Journal of Humanities and Social Science*, 19(4), 99–104.
- Chubb, J., Reed, D., & Cowling, P. (2022). Expert views about missing AI narratives: Is there an AI story crisis? *AI & Society*, 1–20.
- Czarniawska, B. (2009). The uses of narrative in social science research. In M. A. Hardy & A. Bryman (Eds.), *Handbook of Data Analysis* (pp. 649–666). SAGE.
- Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1), 96–138.
- Dogruel, L., Masur, P., & Joeckel, S. (2021). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures*, 1–19.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., ..., & Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71.
- Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, 25(5), 582–602.
- Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Gilster, P. (1997). *Digital literacy*. John Wiley & Sons, Inc.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Greene, T., Shmueli, G., & Ray, S. (forthcoming). Taking the person seriously: Ethically-aware IS research in the era of reinforcement learning-based personalization. *JAIS Preprints (Journal of the Association for Information Systems)*.
- Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3).
- Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., & Karahalios, K. (2021). Attitudes surrounding an imperfect AI autograder. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Jordan, M. I. (2019). Artificial intelligence—the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1).
- Koukouviniou, P., & Holmström, J. (2022). AI Management management beyond the narratives of dystopian nightmares and utopian dreams: a systematic review and synthesis of the literature. In *Proceedings of the European Conference on Information Systems (ECIS)*.
- Kuzior, A., & Kwilinski, A. (2022). Cognitive technologies and artificial intelligence in social perception. *Management Systems in Production Engineering*, 30(2), 109–115.
- Laï, M.-C., Brian, M., & Mamzer, M.-F. (2020). Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France. *Journal of Translational Medicine*, 18(1), 14.
- Leander, K. M., & Burris, S. K. (2020). Critical literacy for a posthuman world: When people read, and become, with machines. *British Journal of Educational Technology*, 51(4), 1262–1276.
- Leiner, D. J. (2014). Our research's breadth lives on convenience samples. A case study of the online respondent pool “SoSci Panel”. *Studies in Communication I Media (SCM)*, 5(4), 367–396.
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the Conference on Human Computer Interaction (HCI)*.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based digital assistants. *Business & Information Systems Engineering*, 61(4), 535–544.
- Maier, S. B., Jussupow, E., & Heinzl, A. (2019). Good, bad, or both? Measurement of physician's ambivalent attitudes towards AI. In *Proceedings of the European Conference on Information Systems (ECIS)*.
- Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution.
- McLean, K. (2013). Literacy and technology in the early years of education: Looking to the familiar to inform educator practice. *Australasian Journal of Early Childhood*, 38(4), 30–41.
- Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (376), 555–572.
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2.
- Pinski, M., Adam, M., & Benlian, A. (2023). AI knowledge: Improving AI delegation through human enablement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Pinski, M., & Benlian, A. (2023). AI literacy - towards measuring human competency in artificial intelligence. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.
- Rai, A., Constantinides, P., & Sarker, S [S.] (2019). Next generation digital platforms toward human-AI hybrids. *MIS Quarterly*, 43(1), iii–ix.
- Sarker, S [Suprateek], Xiao, X., Beaulieu, T., & Lee, A. S. (2018). Learning from first-generation qualitative approaches in the IS discipline: An evolutionary view and some implications for authors and evaluators (PART 1/2). *Journal of the Association for Information Systems*, 19, 752–774.
- Schiff, B. (2012). The function of narrative: Toward a narrative psychology of meaning. *Narrative Matters*, 2(1), 33–47.
- Tapashi, D. (2018). A study on aspects of India's digital divide. *Research Review Journals*, 03(11), 689–693.
- Treiblmaier, H. (2011). Benefits from using continuous rating scales in online survey research. In *Proceedings of the International Conference on Information Systems (ICIS)*.
- Weiss, G. (1999). Multiagent systems: a modern approach to distributed artificial intelligence. *MIT Press*.
- Willcocks, L. (2020). Robo-Apocalypse cancelled? Reframing the automation and future of work debate. *Journal of Information Technology*, 35(4), 286–302.

Paper IV

Hammerschmidt, T.; Stolz, K. & Posegga, O. (2024). *How Leaders' Ambidexterity and Literacy Impact AI Capabilities of Organizations*. Proceedings of 32nd the European Conference on Information Systems (ECIS), Paphos, Cyprus, https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14/.

HOW LEADERS' AMBIDEXTERITY AND LITERACY IMPACT THE AI CAPABILITIES OF ORGANIZATIONS

Completed Research Paper – Accepted Version

Teresa Hammerschmidt, University of Bamberg, Bamberg, Germany, teresa.heyder@uni-bamberg.de

Katharina Stolz, University of Stuttgart, Stuttgart, Germany, katharina.stolz@bwi.uni-stuttgart.de

Oliver Posegga, University of Bamberg, Bamberg, Germany, oliver.posegga@uni-bamberg.de

Abstract

1. *Organizations must align their strategies and renew their capabilities to create business value from advances in artificial intelligence (AI). Ambidextrous leadership is relevant here, as leaders must decide how, when, and where to deploy AI, balancing between its use for increased efficiency (exploitation) and for innovation (exploration). Making such strategic choices calls for AI literacy, comprising certain skills and knowledge. Through an online survey, we identified that leaders' AI literacy is a strong predictor of achieving ambidextrous leadership and making profound choices about which organizational capabilities should be developed to achieve AI transformation successfully. Our findings demonstrate the importance of balancing investments in both tangible organizational resources, especially data governance, and those that are intangible, such as having an open culture with dynamic workforce capabilities. Notably, leaders' AI knowledge is more important than their AI experience for making balanced AI-related decisions.*

2.

Keywords: IT-business alignment, AI literacy, AI capabilities, Ambidexterity

Introduction

Advances in artificial intelligence (AI) offer great potential for organizations to integrate a variety of services accessible to a wide range of customers (Wimelius et al., 2021; Tai et al., 2019; Dwivedi et al., 2023; Radford et al., 2018). Whether organizations can create business value from AI advances depends on how those technologies are aligned with their organizational capabilities and business strategies (Wimelius et al., 2021; Enholm et al., 2022; Dwivedi et al., 2023; Tai et al., 2019). Thus, to benefit from AI progress, organizations must identify and develop suitable capabilities (van de Wetering, 2022; Tai et al., 2019; Vial, 2019; Wimelius et al., 2021; Teubner and Stockhinger, 2020) since the increased autonomy and complexity of AI technologies lead to new forms of “human-machine hybrid work” (Mollick, 2022, p. 5; Dwivedi et al., 2023). This requires new organizational capabilities that are unlike those needed to work with earlier technologies (Mikalef and Gupta, 2021).

To determine which organizational capabilities are required to deal with AI advances successfully, leaders must prioritize AI initiatives and decide when, where, and how AI should be implemented (Alsheibani et al., 2020b; Kolbjørnsrud et al., 2017; Brock and Wangenheim, 2019). In this context, *ambidextrous leadership* is crucial to ensure that decisions made by leaders achieve a balance between the need to leverage the benefits of AI-based automation to optimize existing organizational processes or greater efficiency (i.e., for exploitation) and the need to use AI to reveal opportunities for business innovations and preparing for potential disruption (i.e., for exploration) (Chen et al., 2022; Wirtz, 2020; Kar and Gupta, 2020). Thus, achieving ambidextrous leadership is crucial to benefit from AI transformation as leaders play a crucial role in strategically deploying AI.

Since ambidextrous leadership is based on prioritizing between several options, the outcome of ambidextrous decisions may depend on leaders' available AI-related knowledge and experience as decision makers (i.e., to their *AI literacy*; Long and Magerko, 2020). AI literacy allows for making decisions more accurately and knowledgeably (Chen et al., 2020). However, leaders do not always

possess the skills and knowledge needed to make appropriate decisions about AI that balance exploration and exploitation (Anton et al., 2020; Kolbjørnsrud et al., 2017). This can lead to inadequate decisions at the organizational level and run counter to benefiting from AI advances (Gierlich-Joas et al., 2020). Hence, fostering AI literacy of these decision-makers may be crucial for achieving the ambidextrous leadership required to manage AI transformation successfully.

Generally, little is known about how the capabilities of leaders might affect AI transformation of decisions in organizations (Pinski and Benlian, 2023; Mikalef and Gupta, 2021; van de Wetering, 2022). Specifically, there is a need for more empirical research on what drives leaders' ambidexterity (Wirtz, 2020), in particular, how different degrees of AI literacy within different organizational roles may impact organizational dynamics and the ability to make appropriate decisions about technological transformation (Pinski and Benlian, 2023; Carlsson et al., 2021). To better understand how the AI literacy of leaders affects their ability to make strong, knowledgeable AI-related decisions about their organizations' AI strategies, we ask the following research question: *How is ambidextrous leadership affected by leaders' AI literacy?*

To answer this question, we conducted a quantitative online survey in March–April 2023 of 216 leaders of German-speaking organizations. The survey includes leaders at several levels to ensure a holistic view of organizational decision strategies (Brendel et al., 2021).³ Our survey results yield a tested research model of how leaders' AI literacy impacts ambidextrous leadership and, thus, the development of organizational capabilities in times of AI transformation. Our results provide empirical insights into the relationship between individual and organizational ambidexterity and the interrelationship between leaders' AI literacy and their ability to balance AI-related decisions.

Theoretical Background

In this section, we review the theoretical background on ambidexterity, organizational capabilities, and AI literacy as the main variables for our empirical investigation to answer our research question.

Ambidextrous leadership and organizational ambidexterity

March (1991) described ambidexterity as making tradeoffs between decisions in two different areas: exploration, comprising “search, variation, risk-taking, experimentation ...” and exploitation, comprising “refinement, choice, production, efficiency ...” (p. 71). It is important to balance these two areas, as too much focus on exploration can lead to costs for experimentation without reaping its benefits, whereas too much focus on exploitation can hinder innovation (O'Reilly and Tushman, 2004). Given limited organizational resources, achieving this balance is a significant challenge for leaders, especially in times of technological transformation through service innovations (Rogan and Mors, 2014; Mom et al., 2019; Wolf, 2019).

The literature discusses different types of ambidexterity. Our work focuses on *contextual ambidexterity* as the “choice between alignment-oriented and adoption-oriented activities” (p. 49) of switching between exploration and exploitation at the level of the individual (Birkinshaw and Gibson, 2004; Laureiro-Martínez et al., 2015). Thus, it is about an individual leader's decision-making context. We, thereby, might be aware that leaders within organizations go through temporal cyclings of *sequential ambidexterity*, with long periods of exploitation and short periods of exploration, during which the balance between the two depends on the particular point in time (Chen et al., 2022; Gupta et al., 2006).

To examine how ambidextrous leadership (*leaders' ambidexterity*) at the individual level impacts capabilities to deal with AI at an organizational level, we must understand that leaders' ambidextrous decisions also affect exploitative and explorative choices with respect to Information Technology (IT)-related decisions at that organizational level, such as whether to adopt AI (*organizational ambidexterity*, also known as IT ambidexterity) (Gupta et al., 2006; Kar and Gupta, 2020; O'Reilly and Tushman, 2011). Given that companies that fail to respond appropriately to AI-based changes will miss market

³ Following Kolbjørnsrud et al. (2017), we distinguished between top managers (e.g., CEOs), middle managers (e.g., senior managers), and frontline leaders. Please note that we considered all leadership levels in our work (including managers); we mostly use the term “leaders,” but use “managers” when referring specifically to those in that role in organizations.

opportunities and may even disappear from the market altogether (Chan et al., 2019), many managers today have to make ambidextrous decisions in the face of disruptive digital innovations (Wolf, 2019). Hence, leaders must decide which AI capabilities should be developed (Tai et al., 2019; Vial, 2019; Wimelius et al., 2021; Teubner and Stockhinger, 2020) so their organizations can benefit from AI progress. This represents a duality of decisions between individual and organizational ambidexterity (Gupta et al., 2006; Kar and Gupta, 2020; O'Reilly and Tushman, 2011) that we must consider when seeking to analyze how ambidextrous leadership impacts organizational AI capabilities.

In the current period of widespread use of generative AI leading to new forms of “human-machine hybrid work” (Mollick, 2022, p. 5), ambidextrous leadership requires the development of AI capabilities at both an individual (see Section 2.2) and organizational (see Section 2.3) level (Mikalef and Gupta, 2021; van de Wetering, 2022).

Leaders' AI literacy for appropriate decision making

To strike a balance between the need to improve the efficiency of existing workflows using AI and the need to react to AI disruption with innovations, leaders require specific AI capabilities, known as AI literacy (e.g., Heyder and Posegga, 2021, Long and Magerko, 2020). Not being AI literate can result in bad decisions with respect to how AI is deployed in an organization (Chen et al., 2020; Pinski and Benlian, 2023) and the failure to develop appropriate organizational capabilities, which in turn can limit an organization's ability to benefit from AI transformation (Wimelius et al., 2021; Enholm et al., 2022; Vial, 2019).

We use “AI literacy” to refer to the capabilities of leaders at the individual level. Research has already examined the different capabilities required to enable individuals to deal with AI successfully (e.g., Heyder and Posegga, 2021, Long and Magerko, 2020). The literature distinguishes between different aspects of AI literacy. In the present work, we use the recently published conceptualization of Pinski and Benlian (2023), who distinguish between AI-related knowledge and experience. Knowledge in this context includes the *AI technology knowledge* that makes individuals able to characterize AI technology; *human actors in AI knowledge*, which concerns the role of humans interacting with AI; and *AI steps knowledge*, which refers to specific knowledge of AI inputs, AI processing, and AI output. Experience includes *AI usage experience* and *AI design experience*.

When investigating leaders' AI literacy, we must be aware that there can be interdependencies between AI literacy on the individual level and AI capabilities at the organizational level (see Section 2.3).

Organizational dynamic capabilities

Following Helfat and Peteraf (2003), organizations undergo a continual transformation of their capabilities, which is known as the capability lifecycle and is based on the resource-based theory of firms. Dynamic capabilities are based on organizational learning processes that enable organizations to adapt their strategic decision-making and resources within dynamic markets (Eisenhardt and Martin, 2000). The dynamic capabilities of organizations flow from their leaders' individual capabilities to make sense of technological processes that impact the organizations and address how to employ them for the creation of business value (Teece, 2007). This illustrates an interdependency between individual-level and organizational-level capabilities. However, many organizations struggle to develop appropriate AI capabilities (e.g., Khan et al., 2023; Sjödin et al., 2021). With advances in AI and its increasing self-learning abilities, autonomy, and complexity, a novel capability lifecycle for AI capabilities that are unlike those needed to work successfully with earlier technologies has emerged (Dwivedi et al., 2023; Mikalef and Gupta, 2021; Vial, 2019; Carlsson et al., 2021).

We refer to organizational AI capabilities when discussing firms' dynamic capabilities to deal with AI. Mikalef and Gupta (2021) define AI capabilities at an organizational level as a firm's ability “to select, orchestrate, and leverage its AI-specific resources” (p. 2). They identify three categories of critical organizational capabilities required in the context of AI-based transformation: tangible, human, and intangible. Tangible resources include *data (management) resources*, such as labeled data to train AI systems; the underlying *technological infrastructure*; and *time and financial resources* as basic resources to establish AI projects. Human resources of the workforce include the *technical skills* of all organizational personnel to use AI technology and the *business skills* to use AI in a targeted manner to

derive business value. Following the idea of Teece (2007), this also comprises the dynamic capabilities of individuals to adapt to technological transformation. Intangible resources are an organization's *interdepartmental coordination*, which involves adaptable collaboration among different departments; the *change capacity* to prepare the workforce for changes; and an organization's *risk proclivity* to take some risks by trying something new. Hence, adopting organizational strategies to benefit from transformative changes, such as through AI advances, requires identifying and developing novel organizational dynamic capabilities driven mainly by leaders' individual abilities to recognize the underlying need (Teece, 2007; Carlsson et al., 2021).

Hypotheses and Research Model

In this section, we derive our hypotheses and, in Figure 1, present our multidimensional research model.

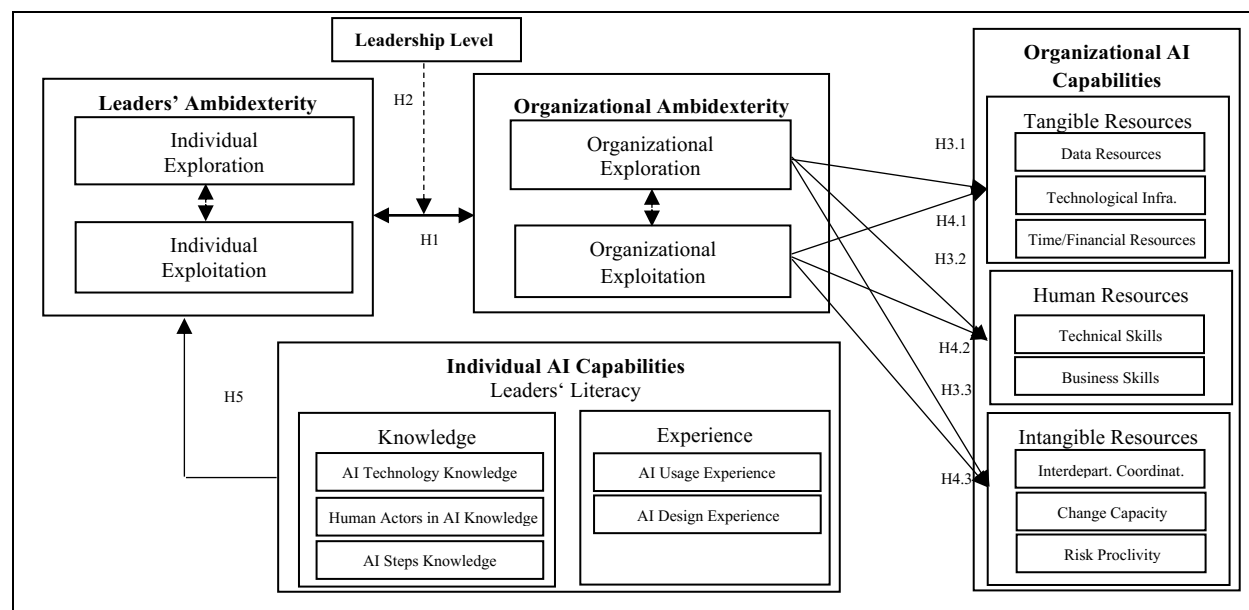


Figure 1. Research model.

Following Chen et al. (2022), individual ambidexterity can affect organizational ambidexterity. Research has already explored how leaders balancing exploration and exploitation at the individual level (*leaders' ambidexterity*) impacts ambidextrous decisions at an organizational level (*organizational ambidexterity*) (Gupta et al., 2006; Kar and Gupta, 2020; O'Reilly and Tushman, 2011). As van de Wetering (2022) argues, ambidextrous leadership is especially relevant when it comes to AI use for innovation (exploration) and routine use (exploitation). Thus, we assume a similar individual level-organizational level interrelationship when it comes to decisions in the context of AI-based transformation, and make the following primary hypotheses. In our hypotheses, organizational ambidexterity acts as a mediator between leaders' ambidexterity as our main independent variable and organizational AI capabilities as our main dependent variable.

H1: *The better leaders balance exploration and exploitation (leaders' ambidexterity), the higher the organization's ability to balance exploration and exploitation (organizational ambidexterity).*

Support for AI closely correlates with a leader's rank in the organization (Kolbjørnsrud et al., 2017). The leader's knowledge of AI is especially crucial, as their interpretation of AI affects its adoption in organizations (Brock and Wangenheim, 2019). Managers tend to be responsible for decisions to implement AI as opposed to working operationally with AI systems (Kolbjørnsrud et al., 2017), so they require managerial knowledge more than technical knowledge (Anton et al., 2020). Of course, this may depend on the management level, as frontline managers have experience with AI that is different than those of top managers (Kolbjørnsrud et al., 2017). There may also be a greater demand for governance-related competencies compared to the workforce, as dealing with ethical issues is a relevant managerial task associated with AI deployment (Anton et al., 2020). Thus, the relationship between individual (*leaders' ambidexterity*) and organizational ambidexterity may be moderated by the *leadership level*. Hence, we assume:

H2: *The relationship between leaders' ambidexterity and organizational ambidexterity is moderated by the leadership level.*

Given the mediating role of *organizational ambidexterity* (comprising organizational exploration and exploitation), we derive several hypotheses characterizing the mediating influence in H3.1 to H3.3 for organizational exploration and H4.1 to H4.3 for organizational exploitation.

When looking at *organizational exploration*, we identify a possible interrelationship between exploration decisions and the development of organizational AI capabilities, which we illustrate as follows. In the context of digital innovation, Holotiuk et al. (2019) argue that “exploration enables (digital) innovation when digital technologies are incorporated” (p. 1). AI innovation requires a renewed technological infrastructure for AI integration (*technological infrastructure*) (Benbya et al., 2021; Jöhnk et al., 2021). Furthermore, given that AI systems are trained rather than programmed, they require large quantities of labeled data (*data resources*) to accomplish sophisticated tasks (Radford et al., 2018). Therefore, appropriate data is required to develop AI-based algorithms (Sjödin et al., 2021). In addition, organizations must invest time and money (*financial resources*) to enhance firm value through the use of AI for novel business opportunities (Lui et al., 2022). Jöhnk et al. (2021) identified that IT infrastructure integration (*technological infrastructure*), the availability of personnel (*time*), and the financial budget (*financial resources*) are factors that demonstrate organizational readiness in the AI adoption process. Therefore, we assume the following hypothesis:

H3.1: *The greater the organizational exploration, the higher the level to which an organization's tangible resources (data resources, technological infrastructure, and time/financial resources) are developed.*

To benefit from the “disruptive and innovative character [of AI] provid[ing] new opportunities” for organizations (*organizational exploration*), AI-related competencies of the workforce become relevant for successful AI adoption (Anton et al., 2020). AI offers some challenges for organizations that differ from earlier technologies, as AI “shifts the locus of action, choice, control, and power away from the exclusive domain of humans” (Benbya et al., 2021, p. 2); thus, in addition to *technical skills* to solve business problems, there is a need for *managerial skills* to identify business problems and ethical issues that arise through the increased autonomy of AI (Anton et al., 2020). Dwivedi et al. (2023) argue that the greater interdependency of generative AI technologies such as ChatGPT requires other skills and a rethinking of the distribution of responsibilities, especially when things do not work out as planned, and so knowledge and exposure to issues in the area of financial and legal liability and responsibility attribution are needed. Thus, we assume:

H3.2: *The greater the organizational exploration, the higher the level to which an organization's human resources (technical skills and business skills) are developed.*

Awareness of AI's potential could contribute to digital innovation strategies by reinforcing creativity and breakthrough behaviors so that innovative products and services can be created (*risk proclivity*) (Hughes, 2018). This demonstrates the characteristics of exploration decision choices (*organizational exploration*) (O'Reilly and Tushman, 2004). New issues arise within the context of generative AI in particular, such as the potential for legal liabilities; these must be identified and considered (Dwivedi et al., 2023). Recent developments in generative AI are also leading to new forms of human-AI collaboration (Mollick, 2022). Organizations must, therefore, develop “agile organizational structures [to be] more flexible” to enhance collaboration among different departments (*interdepartmental coordination*) (Enholm et al. 2022, p. 1727). To be ready for AI-based transformation, they must significantly change their management practices and rethink the way they operate by adopting new business strategies and changes in the workforce's culture (*change capacity*) (Alsheibani et al., 2018; Alsheibani et al., 2020a). Therefore, we assume:

H3.3: *The greater the organizational exploration, the higher the level to which an organization's intangible resources (interdepartmental coordination, change capacity, and risk proclivity) are developed.*

The literature also demonstrates possible interrelationships between exploitation decisions (*organizational exploitation*) and the development of organizational AI capabilities. Given that exploitation at an organizational level is about improving existing IT resources and practices (Chen et

al., 2022), organizations must adopt existing technological assets (*technological infrastructure*), invest in personnel, and allocate financial budget (*time/financial resources*) (Jöhnk et al., 2021) when aiming to use AI to increase the potential for automating existing business processes (Rai et al., 2019). Thus, physical, tangible resources (e.g., *data resources*) are basic requirements (Alvarez-Melgarejo and Torres-Barreto, 2022) to exploit knowledge. Hence, we derive the following hypothesis:

H4.1: *The greater the organizational exploitation, the higher the level to which an organization's tangible resources (data resources, technological infrastructure, and time/financial resources) are developed.*

A firm's *organizational exploitation* is accompanied by learning processes and knowledge (Teubner and Stockhinger, 2020), and companies must reskill their workforces (Kar and Gupta, 2020). Recent service applications based on generative AI offer great potential to increase the efficiency of existing knowledge acquisition practices (Baidoo-Anu and Owusu Ansah, 2023). Identifying AI's optimization potential requires skills and knowledge, such as managerial adaptability and business acumen (*business skills*) (Alsheibani et al., 2020b; Watson et al., 2021; Anton et al., 2020). Employees, too, must develop *technical skills* to promote organizational transformation (Sousa and Rocha, 2019). Hence:

H4.2: *The greater the organizational exploitation, the higher the level to which an organization's human resources (technical skills and business skills) are developed.*

When aiming to automate existing workflows through AI (*organizational exploitation*), organizations must coordinate these automation tools within the organizations' structure, redefine roles, and communicate the underlying changes that come with the integration of AI to optimize business processes (*interdepartmental coordination*) (Benbya et al., 2021). Change management (*change capacity*) will be required; "AI compatibility with the internal organization structure and current processes" is vital to manage organizational changes through the introduction of AI (Alsheibani et al., 2020a, p. 7). Dwivedi et al. (2023) emphasize the importance of having a mindset geared toward a cultural shift aimed at continuous and experimental learning (*risk proclivity*), data orientation, and critical reflection. Hence:

H4.3: *The greater the organizational exploitation, the higher the level to which an organization's intangible resources (interdepartmental coordination, change capacity, and risk proclivity) are developed.*

According to Hinsén et al. (2019), ambidexterity is a critical capability to ensure business survival in hypercompetitive business environments. Leaders can achieve genuine ambidexterity only if they possess the appropriate skills (O'Reilly and Tushman 2004). Absent those skills, decision makers will face problems when seeking to balance the use of AI for innovation (exploration) and optimization (exploitation) (Kar and Gupta, 2020). Furthermore, individual literacy has to be developed; as Laureiro-Martínez et al. (2015) found, exploration and exploitation are accompanied by learning processes (*experience*) and knowledge (*knowledge*). Teubner and Stockhinger (2020) associate the learning mechanisms with *individual exploration* and *individual exploitation*. The learning process of individuals can enrich their capabilities to critically evaluate AI-based technologies and their impact on AI implementation decision making in an organization. Thus, we assume:

H5: *Leaders' literacy (knowledge and experience) positively impacts their ability to manage a balance between exploration and exploitation.*

Method

In this section, we provide an overview of our questionnaire, its operationalization, data collection, data quality, and data analysis – all comprising the quantitative aspects of our empirical investigation.

Questionnaire & operationalization: Our survey employed close-ended items with well-established and empirically proven scales from previous research. The final survey instrument had five sections: sociodemographic information, leadership levels, and experience; leaders' ambidexterity, based on the exploration and exploitation scales of Sirén et al. (2012); organizational ambidexterity, based on the exploration and exploitation scales of Chen et al. (2022) building upon the work of Lee et al. (2015); leaders' AI literacy, based on the operationalization of Pinski and Benlian (2023); and organizational

AI capabilities, based on the operationalization of Mikalef and Gupta (2021). We used sociodemographic data (educational level, age, gender) as control variables.

We excluded participants who did not pass our two aptitude checks assessing their leadership experience (i.e., we excluded participants who did not yet possess any leadership experience) and their experience with AI (we excluded participants who had not yet had at least some AI experience) before they began the first section of the survey. After that, we used mandatory fields, adding an “I don’t know” option to avoid participants making guesses. To reduce the impact of noise in online surveys (Treiblmaier, 2011), we provided a continuous rating scale that measured all quantitative constructs on a five-point Likert scale, with “1” representing “strongly disagree” and “5” representing “strongly agree.” We proved the operationalization by measuring its reliability, following Cronbach’s Alpha as suggested by Boudreau et al. (2001) and Petter et al. (2007). All scales reached the critical level of 0.7 reliability for internal constancy (Cho and Kim, 2015). As our research model in Figure 1 illustrates, all constructs are formative in the third-order dimensions, comprising formative subconstructs in the second-order dimension (representing the scales mentioned above) and reflective indicators in the third-order dimension (representing the items of the scales mentioned above). Following Petter et al. (2007), we used a step-by-step approach to specify the model and represent the relationships between constructs by calculating the composite measures.

Data collection & sample: To reduce the random variation of participants as one sampling error (Brace, 2018), we chose several online channels to acquire participants from different industries, technological levels, and management levels. Leaders were acquired in March–April 2023 through social media, e-mail, and organizational contacts. In particular, LinkedIn was used to reach leaders of German-speaking organizations through the platforms’ search mechanisms and to share the survey link across different LinkedIn groups to promote participation. After potential participants expressed interest via private messaging, we sent them the survey link via e-mail and encouraged them to share the link with their executive contacts. Power analysis showed that 165 participants are required as the effective sample size to achieve .90 power in determining medium-size effect ($F=.15$) of our research model on a significance level of $\alpha=.01$ (Cohen, 1988). Of the 248 participants in total who answered the questionnaire, 216 were usable (21 participants did not pass the aptitude checks at the beginning and 11 participants did not pass the attention check). Given that our sample of leaders was acquired randomly through diverse online channels, we have a convenience sample, following Leiner (2014). As mentioned in the introduction, we included several leadership levels in accordance with Kolbjørnsrud et al. (2017), which comprise top management (11.57%), middle management (21.76%), frontline leaders (48.61%), part-time leaders (0.93%), and other leadership levels (17.13%). Experience ranged between less than 2 years (13.89%), between 2 and 5 years (40.74%), between 6 and 10 years (25.46%), between 11 and 20 years (15.74%), and more than 20 years (4.17%). Most of the top managers are founders of start-ups (73.15%), while most of the middle- and frontline managers are employed leaders within global organizations (80.21%). Males comprise 60.19% of the participants, 39.35% are female, and 0.46% indicated “other” or “prefer not to answer”. Most of our German-speaking leaders had a master’s degree (38.43%) and worked in the information technology (IT) (14.35%), banking/insurance (8.80%), the media (7.87%) industries, or as consultants (10.19%).

Data quality: We followed the validation guidelines for IS (Information Systems) positivist research by Straub and Gefen (2004) and the quality of construct specification of Petter et al. (2007). As already explained, the research model in Figure 1 is multidimensional. Hence, we have several formative constructs in the third-order dimension (e.g., leaders’ literacy, organizational AI capabilities) that build upon several formative constructs in the second-order dimension (e.g., literacy knowledge or tangible resources), with reflective indicators in the first-order dimension (e.g., AI technology knowledge, data). Hence, for testing data quality, we focused on the underlying scales of the first-order dimension. Table 1 is an overview of the reliability and multicollinearity quality tests. To increase content validity, we used only well-established scales that have already been proven empirically and checked the understandability of our questionnaire further with a pre-test of 21 participants (Boudreau et al., 2001). For nomological validity, we relied only on significant values in our analysis (Straub and Gefen, 2004). We applied a correlation matrix for discriminant validity to identify possible multicollinearity, as we have more than one predictor variable in our research model (Field, 2009, p. 224). With the following exceptions between the following first-order constructs, no mean correlations achieved the 0.70 level (see Table 1): *AI technical knowledge* (AITK) and *human actor AI knowledge* (HAK) belonging to

literacy knowledge, *AI usage experience* (AIUE) and *AI design experience* (AIDE) belonging to literacy experience, and *data resources* (ODa), technological infrastructure (OTec), and *other resources* (ORes) belonging to an organizations' tangible resources), as well as *technical skills* (OTS) and *business skills* (OBS) belonging to human resources. Hence, discriminant validity can only be predominantly assumed. However, due to the high correlation results between the literacy scales in the first-order dimension (knowledge: AITK, HAK, and AISK and experience: AIUE and AIDE), we used their underlying higher-order constructs (LiKnow and LiExp) for further analysis. We did so similarly with tangible resources (TangRes) and human resources (HuRes). We further tested the variance inflation factor (VIF) as an alternative measure for testing discriminant validity. By having VIF=2.353 as the highest value, we did not strive for the critical 3.30 level, and thus, multicollinearity is not an issue (Petter et al., 2007).

We evaluated the reliability of our scales by using the standardized values of Cronbach's Alpha (α) as suggested by Boudreau et al. (2001) and Petter et al. (2007). We interpreted the values following Cho and Kim (2015), according to whom all scales can be assumed as reliable if the alpha passes the 0.7 rule. Given that the removal of single items could not significantly improve reliability, we did not remove items from the scales. For the two-item scales, we tested split-half reliability using the Spearman-Brown coefficient with standardized alpha (Eisinga et al., 2013). See Table 1 for an overview:

Constructs	IPlor	IPloi	OPlor	OPloi	AITK	HAK	AISK	AIUE	AIDE	ODa	OTec	ORes	OTS	OBS	OIDC	OCC	ORP
Multicollinearity Tests: Correlation Results																	
Leaders' Ambidexterity (3 rd Order Dimension)																	
IPlor: Ind. Exploration	-																
IPloi: Ind. Exploitation	.660**	-															
Organization's Ambidexterity (3 rd Order Dimension)																	
OPlor: Organ. Explor.	.456**	.419**	-														
OPloi: Organ. Exploit.	.309**	.387**	.459**	-													
Leaders' Literacy (3 rd Order Dimension): Literacy Knowledge (LiKnow; 2 nd Order Dimension)																	
AITK: AI Techn. Knowl.	.405**	.442**	.525**	.292**	-												
HAK: Human AI Knowl.	.373**	.393**	.442**	.313**	.721**	-											
AISK: AI Steps Knowl.	.494**	.485**	.527**	.306**	.762**	.699**	-										
Leaders' Literacy (3 rd Order Dimension): Literacy Experience (LiExp; 2 nd Order Dimension)																	
AIUE: AI Usage Exper.	.337**	.377**	.419**	.177*	.707**	.686**	.669**	-									
AIDE: AI Design Exper.	.281**	.274**	.377**	.121	.792**	.412**	.652**	.532**	-								
Organizational AI Capabilities (3 rd Order Dimension): Tangible Resources (TangRes; 2 nd Order Dimension)																	
ODa: Data Resources	.367**	.434**	.568**	.346**	.548**	.424**	.640**	.432**	.488**	-							
OTec: Technol. Infrastr.	.343**	.368**	.551**	.298**	.502**	.432**	.596**	.462**	.508**	.761**	-						
ORes: Time/Finan. Res.	.343**	.378**	.440**	.252**	.422**	.418**	.580**	.409**	.513**	.735**	.663**	-					
Organizational AI Capabilities (3 rd Order Dimension): Human Resources (HuRes; 2 nd Order Dimension)																	
OTS: Technical Skills	.401**	.388**	.475**	.373**	.523**	.442**	.666**	.450**	.473**	.703**	.672**	.673**	-				
OBS: Business Skills	.431**	.422**	.505**	.254**	.487**	.366**	.588**	.471**	.481**	.684**	.643**	.751**	.766**	-			
Organizational AI Capabilities (3 rd Order Dimension): Intangible Resources (IntanRes; 2 nd Order Dimension)																	
OIDC: Interdep. Coord.	.367**	.456**	.213**	.274**	.247**	.194**	.300**	.244**	.186**	.430**	.321**	.476**	.468**	.567**	-		
OCC: Change Capacity	.419**	.500**	.398**	.347**	.415**	.267**	.388**	.295**	.238**	.543**	.427**	.530**	.592**	.623**	.672**	-	
ORP: Risk Proclivity	.365**	.255**	.297**	.0.77	.214**	.216**	.384**	.307**	.411**	.342**	.390**	.412**	.365**	.420**	.395**	.371**	-
Reliability Tests: Statistics																	
Scale(s) are based on	Sirén et al. (2012)		Chen et al. (2022)		Pinski and Benlian (2023)					Mikalef and Gupta (2021)							
n	216	216	216	216	210	210	210	210	209	203	205	194	206	205	212	213	212
Nbr. of Items	6	6	3	3	3	3	3	2	2	6	7	3	7	7	7	6	3
					LiKnow (9): $\alpha=.932$		LiExp (4): $\alpha=.835$		TangRes (16): $\alpha=.891$		HuRes (14): $\alpha=.867$		IntangRes (16): $\alpha=.736$				
α	.776	.764	.858	.790	.875	.830	.932	.762	.930	.923	.927	.919	.927	.947	.893	.837	.893
Mean Values	3.74	3.84	3.45	3.85	3.52	3.67	3.28	3.45	2.66	3.30	3.21	3.01	3.39	3.30	3.75	3.69	2.91
Stand. Dev.	.599	.598	.935	.745		.834		.979		.952			.886				.666

**significant at the 0.01 level, *significant at the 0.05 level

Table 1. Reliability and correlation results.

Data analysis: We used SPSS to conduct our statistical analysis. We tested the distribution of variables for each scale with Kolmogorov-Smirnov and Shapiro-Wilk (Razali and Wah, 2011). Since the normal distribution is not given, we applied the Spearman coefficient to measure linear correlation (Winter et al., 2016). We used the correlation to test whether there are some interdependencies in the first step (Taylor, 1997, p. 217). In addition, we used correlation to avoid isolated interpretation of regression results and to check multicollinearity further (Courville and Thompson, 2001). We applied multivariate regression to prove the direction and scope of influence, since we have several independent and dependent variables in our research model (Marill, 2004). For H1, H2, and H5, we used the adjusted R²

to show how much of the variation of the dependent variable is based on the independent variable (Marill, 2004). Wilk's Lambda (Λ) was used via MANOVA to measure the error variance between groups (e.g., LiKnow compared to LiExp or IPlor compared to IPloi; see results for details) (Brown and Tinsley, 1983). For the leadership level as a moderating variable (H2), we used ordinary least square (OLS) regression via SPSS AMOS (Hayes and Matthes, 2009; Memon et al., 2019). For organizational ambidexterity as a mediating variable (H3 and H4), we used the Sobel test (Abu-Bader and Jones, 2021). We thus tested the influence between the main variables with the beta coefficient (β) as standardized regression coefficient for linear multiple regression (Courville and Thompson, 2001).

Results

Table 2 presents the empirical results leading to our hypotheses validation.

Hypotheses		Variables	Regression		
H	Validation	IV $\rightarrow$ DV	Coefficients	p	SD/SE
H1	supported	IPlor $\rightarrow$ OPlor	adj. $R^2 = .710$	.001	.095
		IPloi $\rightarrow$ OPloi	adj. $R^2 = .485$	.001	.078
H2	not supported	IPlor $\rightarrow$ Oplor (Moderator: LL*)	$\beta = -.099$	.182	.074
		IPloi $\rightarrow$ OPloi (Moderator: LL*)	$\beta = -.122$	.044	.061
H3.1	supported	IPlor*IPloi $\rightarrow$ (Mediator: OPlor) $\rightarrow$ TangRes	a=.823, b=.478, $S_a = .001$, $S_b = .001$, Z = 413.34	IPlor*IPloi $\rightarrow$ TangRes: adj. $R^2 = .196$, p=.000;	.000
H4.1	supported	IPlor*IPloi $\rightarrow$ (Mediator: OPloi) $\rightarrow$ TangRes	a=.525, b=.217, $S_a = .001$, $S_b = .012$, Z = 18.07	IPlor: $\beta = .200$, p=.016 IPloi: $\beta = .293$, p=.000	.000
H3.2	supported	IPlor*IPloi $\rightarrow$ (Mediator: OPlor) $\rightarrow$ HuRes	a=.823, b=.353, $S_a = .001$, $S_b = .001$, Z = 324.42	IPlor*IPloi $\rightarrow$ HuRes: adj. $R^2 = .208$, p=.000;	.000
H4.2	supported	IPlor*IPloi $\rightarrow$ (Mediator: OPloi) $\rightarrow$ HuRes	a=.525, b=.182, $S_a = .001$, $S_b = .023$, Z = 7.91	IPlor: $\beta = .255$, p=.003 IPloi: $\beta = .254$, p=.003	.000
H3.3	supported	IPlor*IPloi $\rightarrow$ (Mediator: OPlor) $\rightarrow$ IntanRes	a=.823, b=.122, $S_a = .001$, $S_b = .010$, Z = 12.20	IPlor*IPloi $\rightarrow$ IntanRes: adj. $R^2 = .260$, p=.000;	.000
H4.3	not supported	IPlor*IPloi $\rightarrow$ (Mediator: OPloi) $\rightarrow$ IntanRes	a=.525, b=.065, $S_a = .001$, $S_b = .251$, Z = 0.26	IPlor: $\beta = .275$, p=.001 IPloi: $\beta = .291$, p=.000	.080
H5	supported	LiKno $\rightarrow$ IPlor	adj. $R^2 = .330$	.001	.043
		LiExp $\rightarrow$ IPloi	adj. $R^2 = .213$	.001	.038

*IV=Independent Variable, DV=Dependent Variable, adj. R^2 =Adjusted Regression Coefficient, β =Beta Weight, a/b=Unstandardized Regression Coefficients, S_a/S_b =Standard Errors, Z=Sobel Test Statistic, p=Significance, SD=Standard Deviation, SE=Standard Error

Table 2. Regression results.

Regarding our first hypothesis, the significant and high correlation between individual exploration and exploitation demonstrates a balance between both variables, indicating that leaders' ambidexterity can be assumed. Compared to this, organization exploration and exploitation are balanced with only moderate correlation and regression results. Thus, organizational ambidexterity is lower compared to leaders' ambidexterity. One reason might be the high and insignificant error variance of Wilk's Lambda of individual exploration influencing organizational ambidexterity ($\Lambda = .735$, $p = .258$). Hence, individual exploitation better explains the variation of both variables, organization exploration and exploitation ($\Lambda = .561$, $p = .000$). Given that regression results were at a moderate level, individual exploration and exploitation of leaders' ambidexterity can explain a moderate proportion of the variation of organizational ambidexterity; which supports H1, namely that the better leaders balance exploration and exploitation, the higher their expectations regarding the organization's ability to achieve ambidexterity.

Regarding our second hypothesis testing the moderating influence of leadership level on the relationship between leaders' ambidexterity and organizational ambidexterity (H2), results indicate low and significant correlations (see Table 2), except when it comes to organization exploitation ($r = -.048$, $p = .479$). Given that we have a non-significant and high error variance of Wilk's Lambda ($\Lambda = .983$, $p = .396$) and further non-significant regression results, we cannot support H2, meaning that the leadership level does not moderate the relationship between leaders and organizational ambidexterity. This might be because most of our participants reported having 2 to 5 years' work experience (40.74%) or between 6 to 10 years (25.46%), and thus may not have as much specific decision-making experience as leaders with more years of overall work experience.

Regarding our third and fourth hypotheses, when it comes to the influence of organizational ambidexterity and organizational AI capabilities (H3 and H4), we aligned the sub-hypotheses of organizational exploration and exploitation to the subconstructs: tangible, human, and intangible resources to better estimate multivariate influences of organizational ambidexterity according to our underlying research model.

Tangible resources: Beginning with the impact of organizational ambidexterity on tangible resources, we see especially high and significant correlations between organization exploration and data and technology capabilities. In general, correlation results are higher between organization exploration and tangible resources compared to organization exploitation. This is further shown through the better error variance when looking at Wilk's Lambda (organization exploration: $\Lambda=.567$, $p=.000$; organization exploitation: $\Lambda=.654$, $p=.003$). Due to the significant and moderate regression results, we can support H3.1 and H4.1 and, thus, organization exploration better explains variations in tangible resources.

Human resources: Looking at the impact of organizational ambidexterity on human resources, we have significant moderate-high correlation results: the correlation coefficient is higher between organization exploration and human resources compared to organization exploitation (see Table 2). This is supported by the higher error variance of Wilk's Lambda of organization exploitation ($\Lambda=.788$, $p=.067$) compared to organization exploration ($\Lambda=.703$, $p=.002$). However, we have a high error variance in general, indicating no clear separation between technical skills and business skills. Due to the significant and moderate regression results, we support H3.2 and H4.2 – meaning that the greater a leader's expectations regarding organizational exploration and exploitation, the higher the level to which an organization's human resources (technical and business skills) are developed.

Intangible resources: Investigating the impact of organizational ambidexterity on intangible resources, we first had to eliminate risk proclivity due to the non-significant correlation (see Table 1). The regression analysis in Table 2 shows low and significant results for H3.3 but non-significant results for H4.3, with estimates that are too inaccurate given the low z-value. The error variance of organization exploration ($\Lambda=.567$, $p=.000$) is better than organization exploitation ($\Lambda=.654$, $p=.003$). We can support H3.3 (mediator: organizational exploration) but not H4.3 (mediator: organizational exploitation). One reason might be that the change capacity and interdepartmental coordination as organizational resources are by definition closer to transformative decision-making choices (i.e., to exploration).

Regarding our fifth hypothesis, when analyzing the impact of leaders' literacy on human resources, the results indicate significant and high correlation results and moderate-high significant regression results, thus supporting H5. Technical skills, in particular, are affected by leaders' literacy. We further checked the impact of leaders' literacy on leaders' ambidexterity. We have significant and high correlation results and low-moderate and significant regression coefficients. In both cases, the error variance of Wilk's Lambda is significantly lower when having literacy-related knowledge ($\Lambda=.263$, $p=.021$) compared to experience ($\Lambda=.454$, $p=.011$). Hence, knowledge has a greater influence than experience on leaders' ambidexterity.

Control variables: We further tested whether the sociodemographic variables might influence leaders' literacy. We found only a low and significant correlation between knowledge and years of leadership experience ($r=.154^*$, $p=.025$) and a negative, low, and significant correlation between knowledge and leadership level ($r=.174^*$, $p=.012$). Thus, we did not conduct further analyses.

To summarize, we observe that organization exploration better explains the variation of organizational AI capabilities compared to organization exploitation. Individual exploitation better explains the variation of organizational ambidexterity compared to individual exploration and, thus, leaders' literacy (especially their AI knowledge) influences the balance between individual exploitation and exploration.

Discussion

Below, we discuss theoretical and practical implications and discuss our work's limitations.

Theoretical implications: AI knowledge for ambidextrous leadership

We found a greater balance between exploration and exploitation of individuals compared to that of organizations, but this does not negate that ambidexterity can be more easily achieved at the organizational level than by individuals (Gupta et al., 2006) because it is easier for organizations to cycle between “periods of exploration and exploitation” (Chen et al., 2022) and adopt different strategies within different business units (Tai et al., 2019). It can be challenging for individuals to switch between exploration and exploitation, as they are typically responsible for specific operational areas that focus either on one or the other (e.g., being a leader of the research and development department or leading

IT security) (Bidmon and Boe-Lillegraven, 2020). Thus, future work might examine the interdependencies of ambidexterity between the individual and organizational levels in-depth.

Our findings further showed that within AI literacy, knowledge demonstrates a greater positive impact than experience. This can be attributed to the fact that leaders make decisions about AI technologies rather than using them directly (Kolbjørnsrud et al., 2017). Thus, leaders' AI knowledge may be more important than experience in achieving ambidexterity. It would be interesting to compare these results with the ambidexterity of other individuals, such as AI developers or users, given that their roles in human-AI interaction differ from those of managers (Kolbjørnsrud et al., 2017). Future work should examine how different subconstructs of AI literacy might impact ambidextrous decision-making in a different manner by comparing groups of organizational deciders.

Our results further highlight the impact of the decision-making context on ambidextrous leadership since leaders who cannot balance exploration and exploitation are often found in organizations with a low risk proclivity. This is supported by previous literature, which states that the culture of the direct working environment can impact whether ambidexterity can be achieved (e.g., Alsheibani et al., 2018, Alsheibani et al., 2020a). This demonstrates that it is not only the AI capabilities of leaders that impact ambidextrous decisions and, thus, the development of organizational capabilities, but that the reverse is also true: organizational capabilities impact the capabilities of individuals deciding within an organizational context (referring to the nature of dynamic capabilities; see Section 2.3). Hence, traditional resource-based approaches may have become inadequate for explaining whether organizations fail or succeed in implementing AI (e.g., Carlsson et al., 2021; Teece, 2007). Hence, future work on AI adoption might focus more on the idea of a dynamic capability lifecycle that undergoes AI transformation.

Table 3 summarizes future research avenues based on the theoretical implications of our work:

Research Area	Possible Research Question
AI Ambidexterity	How does leaders' ambidexterity form organizational ambidexterity on an aggregate level?
AI literacy	How do AI literacy knowledge and experience contribute to the ability of different groups of individuals (e.g., leaders, developers, workers) to balance exploration and exploitation?
AI adoption	How can we explain the relationship between organizational and leaders' capabilities in times of AI transformation?

Table 3. *Future Research Avenues.*

Practical implications: Fostering dynamic and intangible capabilities

Our results further indicate that among the intangible resources in the conceptualization of Mikalef and Gupta (2021), risk proclivity was rated as the lowest compared to interdepartmental coordination and change management capacity. It did not exhibit a significant relationship with organizations' ambidexterity. This challenges our assumption that the risk-taking abilities of organizations are affected by ambidextrous decision making and suggests that other intangible capabilities, such as ethics and culture (not considered in detail in the Mikalef and Gupta (2021) conceptualization), might be more critical. Ethics and culture have become especially relevant in the increasingly complex use cases and services emerging from recent advances in generative AI (Dwivedi et al., 2023). Leaders, thus, should act as role models, providing guidance for their workforce to foster their collaboration intentions and motivation to continuously develop dynamic capabilities to deal with AI transformation.

Further, our results highlight that organizations lack adequate tangible resources, especially data and technology (following the Mikalef and Gupta (2021) conceptualization). Investments in data governance and technological infrastructure are required to deal with AI advances (Dwivedi et al., 2023). Hence, a variety of tangible, intangible, and human resources are essential for successful AI deployment. However, because organizational resources are limited, organizations must balance their investments. Given the rapid progress of AI innovations and how they change the ways in which humans work with technology (Mollick, 2022; Baird and Maruping, 2021; Maedche et al., 2019), leaders should prioritize more than just the improvement of existing workflows and operations. Following the idea of establishing dynamic capabilities to deal successfully with ongoing AI transformation (Carlsson et al., 2021), leaders need to sensitize the workforce to AI's impacts, especially the new forms of human-AI hybrids (not fully under organizations' control) that emerge from AI's higher level of autonomy and interactivity, the increased interconnectedness of different actors and resources, and the fact that generative AI is easily accessible to everyone (Dwivedi et al., 2023; Maedche et al., 2019; Wolf, 2019; Mollick, 2022).

Hence, organizations should prioritize rather intangible capability development to prepare their individual leaders and workers to dynamically adapt to ongoing AI transformation.

Limitations and future work

Our work has some limitations. *First*, there may be other types of ambidexterity, such as the structural ambidexterity that separates the exploration and exploitation activities of different business units (Tushman and O'Reilly III, 1996), which may mediate the relationship between ambidexterity at the individual and organizational levels. Comparing ambidexterity between different departments and between different leaders would be interesting for future research.

Second, we used the operationalization of earlier work that did not consider responsibility perceptions in detail; thus, our online survey did not include ethical capabilities as part of ambidextrous decision making. However, with the wide accessibility of generative AI with ever-greater autonomy and interconnectedness (Dwivedi et al., 2023), we must ensure that decisions about using these technologies are guided by ethical considerations. Future work should explicitly integrate ethical capabilities when considering individual and organizational ambidexterity in the context of AI advances.

Third, there may be further relationships between individual-level AI literacy and human resources as AI capabilities on the organizational level, given that an initial analysis indicated significant correlations: leaders' AI knowledge tends to be related to technical skills, while AI experience is more related to business skills. Hence, future work might examine this relationship in more detail.

Fourth, there may be differences with respect to capabilities that are perceived to be relevant depending on how leaders are affected by AI transformation, given that our survey took place in the midst of an ongoing public debate regarding generative AI (Mollick, 2022). Future work might also incorporate the perspective of employees, whose evaluation of a firm's dynamic capability lifecycle might be more realistic and critical than that of managers, and to avoid sample biases through public discourse.

Conclusion

This work highlights the importance of ambidextrous decision-making of leaders so they can strike a balance between exploration and exploitation in organizations and thus develop capabilities that ready their organizations for advances in generative AI. Our results support our hypothesis that AI literacy can help leaders make ambidextrous decisions. Within AI literacy, managers' AI knowledge may be more important than their AI experience when it comes to achieving a balance between exploration and exploitation in those AI-related decisions. Hence, leadership training is required to strengthen their AI knowledge. Our results further demonstrate the importance of balancing investments in different organizational resources that are tangible (especially data governance) and intangible (especially an open and ethical-sensitized culture) when aiming to foster beneficial AI strategies. To summarize, increasing the knowledge that enables leaders to make informed and balanced decisions regarding how to deploy and use AI organizations contributes to those leaders' ambidexterity. It reduces stress and procrastination, diminishes the negative consequences that come from imbalanced decisions, and ultimately contributes to the development of adequate organizational capabilities. We hope our work adds new elements to the current debate on how to deal with generative AI by providing insights into how ambidexterity and literacy are interrelated between the individual and organizational levels.

References

- Abu-Bader, S. and T. V. Jones (2021). "Statistical Mediation Analysis Using the Sobel Test and Hayes SPSS Process Macro" *International Journal of Quantitative and Qualitative Research Methods*.
- Alsheibani, S. A., Y. Cheung and C. Messom (2018). "Artificial Intelligence Adoption: AI-readiness at Firm-Level". In *Pacific Asia Conference on Information Systems (PACIS) Proceedings*, 37-44.
- Alsheibani, S. A., Y. Cheung, C. Messom and M. Alhosni (2020a). "Winning AI Strategy: Six-Steps to Create Value from Artificial Intelligence". In *Americas Conference on Information Systems (AMCIS) Proceedings*.

- Alsheibani, S. A., D. Messom, Y. Cheung and M. Alhosni (2020b). "Reimagining the Strategic Management of Artificial Intelligence: Five Recommendations for Business leaders". In *Americas Conference on Information Systems (AMCIS) Proceedings*.
- Alvarez-Melgarejo, M. and M. L. Torres-Barreto (2022). "Tangible Resources and Integration Capabilities as Determinants of Knowledge Creation Capabilities" *Revista EAN* (92).
- Anton, E., A. Behne and F. Teuteberg (2020). "The Humans Behind Artificial Intelligence—An Operationalisation of AI Competencies". In: *European Conference on Information Systems (ECIS) Proceedings*.
- Baidoo-Anu, D. and L. Owusu Ansah (2023). "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning" *Available at SSRN 4337484*.
- Baird, A. and L. M. Maruping (2021). "The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts" *MIS Quarterly* 45 (1), 315–341.
- Benbya, H., S. Pachidi and S. Jarvenpaa (2021). "Special Issue Editorial: Artificial Intelligence in Organizations: Current State and Future Opportunities" *Journal of the Association for Information Systems* 22 (2).
- Bidmon, C. M. and S. Boe-Lillegraven (2020). "Now, switch! Individuals' responses to imposed switches between exploration and exploitation" *Long Range Planning* 53 (6).
- Birkinshaw, J. and C. Gibson (2004). "Building Ambidexterity into an Organisation" *MIT Sloan Management Review*, 47–55.
- Boudreau, M. C., D. Gefen and D. W. Straub (2001). "Validation in Information Systems Research: A State-of-the-Art Assessment" *MIS Quarterly* 25 (1), 1–16.
- Brace, I. (2018). *Questionnaire design. How to plan, structure and write survey material for effective market research*. UK: Kogan Page Publisher.
- Brendel, A. B., M. Mirbabaie, T.-B. Lembcke and L. Hofeditz (2021). "Ethical Management of Artificial Intelligence" *Sustainability* 13 (4), 1974.
- Brock, J. K.-U. and F. von Wangenheim (2019). "Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence" *California Management Review* 61 (4), 110–134.
- Brown, M. T. and H. E. Tinsley (1983). "Discriminant analysis" *Journal of Leisure Research* 15 (4).
- Carlsson, L., M. Hattinger, A. K. Olsson and U. Lundh Snis (2021). "Desperately seeking industrial digital strategy : a dynamic capability approach" *International Journal of Information Systems and Change Management* 12 (4), 345–364.
- Chan, C. M., S. Y. Teoh, A. Yeow and G. Pan (2019). "Agility in responding to disruptive digital innovation: Case study of an SME" *Information Systems Journal* 29 (2), 436–455.
- Chen, Y., H. Liu and M. Chen (2022). "Achieving novelty and efficiency in business model design: Striking a balance between IT exploration and exploitation" *Information & Management* 59 (3).
- Chen, Y., T. Wen and H. Huang (2020). *The Importance of AI Literacy in Decision Making*.
- Cho, E. and S. Kim (2015). "Cronbach's coefficient alpha" *Organizational Research Methods* 18 (2), 207–230.
- Cohen, J. (1988). "Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers" *Computers, Environment and Urban Systems* 14.
- Courville, T. and B. Thompson (2001). "Use of Structure Coefficients in Published Multiple Regression Articles: β is not Enough" *Educational and Psychological Measurement* 61 (2), 229–248.
- Dwivedi, Y. K., N. Kshetri, L. Hughes, ... and R. Wright (2023). "“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy" *International Journal of Information Management* 71.
- Eisenhardt, K. M. and J. A. Martin (2000). "Dynamic Capabilities: What Are They?" *Strategic management journal* 21 (10/11), 1105–1121.
- Eisinga, R., M. t. Grotenhuis and B. Pelzer (2013). "The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?" *International journal of public health* 58 (4), 637–642.
- Enholm, I. M., E. Papagiannidis, P. Mikalef and J. Krogstie (2022). "Artificial Intelligence and Business Value: a Literature Review" *Information Systems Frontiers* 24 (5), 1709–1734.
- Field, A. (2009). *Discovering Statistics Using SPSS. Introducing Statistical Method*: SAGE.

- Gierlich-Joas, M., T. Hess and R. Neuburger (2020). “More self-organization, more control—or even both? Inverse transparency as a digital leadership concept” *Business Research* 13 (3), 921–947.
- Gupta, A. K., K. G. Smith and C. E. Shalley (2006). “The interplay between exploration and exploitation” *Academy of management journal* 49 (4), 693–706.
- Hayes, A. F. and J. Matthes (2009). “Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations” *Behavior Research Methods* 41 (3), 924–936.
- Helfat, C. E. and M. A. Peteraf (2003). “The dynamic resource-based view: capability lifecycles” *Strategic Management Journal* 24 (10), 997–1010.
- Heyder, T. and O. Posegga (2021). “Extending the foundations of AI literacy”. In: *Proceedings of the International Conference on Information Systems (ICIS)*.
- Hinsen, S., J. Jöhnk and N. Urbach (2019). “Disentangling the Concept and Role of Continuous Change for IS Research : A Systematic Literature Review”. In *International Conference on Information Systems (ICIS) Proceedings*.
- Holotiuk, F., Beimborn and Daniel (2019). “Temporal Ambidexterity: How Digital Innovation Labs Connect Exploration and Exploitation for Digital Innovation”. In *International Conference on Information Systems (ICIS) Proceedings*.
- Hughes, M. (2018). “Organisational ambidexterity and firm performance: burning research questions for marketing scholars” *Journal of Marketing Management* 34 (1-2), 178–229.
- Jöhnk, J., M. Weißert and K. Wyrski (2021). “Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors” *Business & Information Systems Engineering* 63 (1), 5–20.
- Kar, A. and M. P. Gupta (2020). “Talent Scarcity, Skill Distance and Reskilling Resistance in Emerging Digital Technologies - Understanding Employee Behaviour”. In *International Conference on Information Systems (ICIS) Proceedings*.
- Khan, A. N., F. Jabeen, K. Mehmood, M. Ali Soomro and S. Bresciani (2023). “Paving the way for technological innovation through adoption of artificial intelligence in conservative industries” *Journal of Business Research* 165 (C).
- Kolbjørnsrud, V., R. Amico and R. J. Thomas (2017). “Partnering with AI: how organizations can win over skeptical managers” *Strategy & Leadership* 45 (1), 37–43.
- Laureiro-Martínez, D., S. Brusoni, N. Canessa and M. Zollo (2015). “Understanding the exploration-exploitation dilemma: An fMRI study of attention control and decision-making performance” *Strategic Management Journal* 36 (3), 319–338.
- Lee, O.-K., V. Sambamurthy, K. H. Lim and K. K. Wei (2015). “How Does IT Ambidexterity Impact Organizational Agility?” *Information Systems Research* 26 (2), 398–417.
- Leiner, D. J. (2014). “Our research's breadth lives on convenience samples. A case study of the online respondent pool “SoSci Panel”” *Studies in Communication I Media (SCM)* 5 (4), 367–396.
- Long, D. and B. Magerko (2020). “What is AI literacy? Competencies and design considerations”. In *Proceedings of the Conference on Human Computer Interaction (HCI)*.
- Lui, A. K., M. C. Lee and E. W. Ngai (2022). “Impact of artificial intelligence investment on firm value” *Annals of Operations Research*, 1–16.
- Maedche, A., C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana and M. Söllner (2019). “AI-Based digital assistants” *Business & Information Systems Engineering* 61 (4), 535–544.
- March, J. G. (1991). “Exploration and Exploitation in Organizational Learning” *Organization Science* 2 (1), 71–87.
- Marill, K. A. (2004). “Advanced Statistics: Linear Regression, Part I: Simple Linear Regression” *Academic Emergency Medicine* 11 (1), 87–93.
- Memon, M. A., J.-H. Cheah, T. Ramayah, H. Ting, F. Chuah and T. H. Cham (2019). “Moderation Analysis: Issues and Guidelines” *Journal of Applied Structural Equation Modeling* 3 (1), i–xi.
- Mikalef, P. and M. Gupta (2021). “Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance” *Information & Management* 58 (3).
- Mollick, E. (2022). “ChatGPT is a tipping point for AI” *Harvard Business Review*.
- Mom, T. J. M., Y.-Y. Chang, M. Cholakova and J. J. P. Jansen (2019). “A Multilevel Integrated Framework of Firm HR Practices, Individual Ambidexterity, and Organizational Ambidexterity” *Journal of Management* 45 (7), 3009–3034.
- O'Reilly, C. A. and M. L. Tushman (2004). “The ambidextrous organization” *Harvard Business Review* 82 (4), 74–81.

- O'Reilly, C. A. and M. L. Tushman (2011). "Organizational Ambidexterity in Action: How Managers Explore and Exploit" *California Management Review* 53 (4), 5–22.
- Petter, S., D. Straub and A. Rai (2007). "Specifying Formative Constructs in Information Systems Research" *MIS Quarterly* 31 (4), 623–656.
- Pinski, M. and A. Benlian (2023). "AI literacy - towards measuring human competency in artificial intelligence". In: *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.
- Radford, A., K. Narasimhan, T. Salimans and I. Sutskever (2018). "Improving language understanding by generative pre-training".
- Rai, A., P. Constantinides and S. Sarker (2019). "Next generation digital platforms toward human-AI hybrids" *MIS Quarterly* 43 (1), iii–ix.
- Razali, N. M. and Y. B. Wah (2011). "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests" *Journal of statistical modeling and analytics* 2 (1), 21–33.
- Rogan, M. and M. L. Mors (2014). "A Network Perspective on Individual-Level Ambidexterity in Organizations" *Organization Science* 25 (6), 1860–1877.
- Sirén, C. A., M. Kohtamäki and A. Kuckertz (2012). "Exploration and exploitation strategies, profit performance, and the mediating role of strategic learning: Escaping the exploitation trap" *Strategic Entrepreneurship Journal* 6 (1), 18–41.
- Sjödin, D., V. Parida, M. Palmié and J. Wincent (2021). "How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops" *Journal of Business Research* 134, 574–587.
- Sousa, M. J. and Á. Rocha (2019). "Digital learning: Developing skills for digital transformation of organizations" *Future Generation Computer Systems* 91, 327–334.
- Straub, D. and D. Gefen (2004). "Validation Guidelines for IS Positivist Research" *Communications of the Association for Information Systems* 13.
- Tai, J. C., E. T. Wang and H.-Y. Yeh (2019). "A study of IS assets, IS ambidexterity, and IS alignment: the dynamic managerial capability perspective" *Information & Management* 56 (1), 55–69.
- Taylor, J. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements*. Harvard: University Science Books.
- Teece, D. J. (2007). "Explicating Dynamic Capabilities: The Nature and Microfoundations of (Sustainable) Enterprise Performance" *Strategic management journal* 28 (13), 1319–1350.
- Teubner, R. A. and J. Stockhinger (2020). "Literature review: Understanding information systems strategy in the digital age" *The Journal of Strategic Information Systems* 29 (4), 101642.
- Treiblmaier, H. (2011). "Benefits from using continuous rating scales in online survey research". In *Proceedings of the International Conference on Information Systems (ICIS)*.
- Tushman, M. L. and C. A. O'Reilly III (1996). "The Ambidextrous Organization: Managing Evolutionary and Revolutionary Change" *California Management Review* 38 (4), 8–30.
- van de Wetering, R. (2022). "The impact of artificial intelligence ambidexterity and strategic flexibility on operational ambidexterity". In *Pacific Asia Conference on Information Systems (PACIS) Proceedings*.
- Vial, G. (2019). "Understanding digital transformation: A review and a research agenda" *The Journal of Strategic Information Systems* 28 (2), 118–144.
- Watson, G. J., K. C. Desouza, V. M. Ribiere and J. Lindič (2021). "Will AI ever sit at the C-suite table? The future of senior leadership" *Business Horizons* 64 (4), 465–474.
- Wimelius, H., L. Mathiassen, J. Holmström and M. Keil (2021). "A paradoxical perspective on technology renewal in digital transformation" *Information Systems Journal* 31 (1), 198–225.
- Winter, J. C. F. de, S. D. Gosling and J. Potter (2016). "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data" *Psychological methods* 21 (3), 273–290.
- Wirtz, J. (2020). "Organizational Ambidexterity: Cost-Effective Service Excellence, Service Robots, and Artificial Intelligence" *Organizational dynamics* 49 (3), 100719.
- Wolf, V. (2019). "Ambidexterity in service innovation research: A systematic literature review". In: *International Conference on Wirtschaftsinformatik (WI) Proceedings*.

Paper V

Heyder, T. & Posegga (2021). *Extending the Foundations of AI Literacy* (Short Paper, Best Paper Nominee).
Proceedings of the 46nd International Conference on Information Systems (ICIS), Austin, Texas,
https://aisel.aisnet.org/icis2021/is_future_work/is_future_work/9/.

Extending the foundations of AI literacy

Short Paper – Accepted Version

Teresa Heyder

University of Bamberg
An der Weberei 5, 96047 Bamberg
teresa.heyder@uni-bamberg.de

Oliver Posegga

University of Bamberg
An der Weberei 5, 96047 Bamberg
oliver.posegga@uni-bamberg.de

Abstract

Artificial Intelligence (AI) has the potential to change the future of work by improving employee task performance. Unlocking this potential requires that organizations prioritize developing employee AI competencies. Our understanding of AI competencies required to engage with AI-based systems effectively, however, is limited. The concept of AI literacy serves as a starting point to map and organize the competencies needed to use, explore, and communicate (with) AI. This short paper extends previous conceptualizations of AI literacy by exploring its dimensions and characteristics in organizational settings, both conceptually and empirically. We present insights into a multimethod taxonomy development process and a preliminary taxonomy of AI literacy. Our results strengthen and extend the conceptualization of the functional, critical, and sociocultural dimensions of AI literacy. We identify potential relationships between these dimensions and find that sociocultural AI literacy plays a significant role in enabling employees to engage with AI effectively at work.

Keywords: Artificial Intelligence, literacy taxonomy, AI literacy, augmentation, future skills

Introduction

Following Rai et al. (2019, p. iii), AI can be defined as “the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning,” and so on. The most recent generation of AI-based systems, such as voice- and text-based assistants (Maedche et al. 2019), changes the nature of the symbiosis between humans and technology and increases the capabilities of both: AI enables humans to perform tasks better, and humans augment AI through training and judgment (Rai et al. 2019). AI-based technologies, such as the most recent generation of intelligent assistants, will change work by “augmenting human task performance with a higher extent of interactivity and intelligence than previous generations of digital assistants or traditional software applications” (Maedche et al. 2019, p. 536). Augmentation, the use of (AI-based) technologies to improve human performance and collaboration, poses novel management challenges (Rai et al. 2019). Understanding and developing the competencies of employees working with these technologies is a major challenge (Hattinger and Eriksson 2020). A limited understanding of AI and a lack of the required competencies to engage with the technology effectively have a negative effect on AI use (Long and Magerko 2020), leading to misconceptions about or misuse of the technology.

In this short paper, we present the results of an ongoing research project aimed at understanding how AI-based technologies augment human actors, and vice versa. More specifically, we are interested in the role AI literacy plays in the augmentation process.

Long and Magerko (2020, p. 1) define *AI literacy* as a set of competencies individuals need to use AI successfully and that transform how humans communicate and collaborate with AI technologies. AI literacy offers a broad concept of competencies to use, explore, and communicate (with) AI at home and in the workplace (Long and Magerko 2020, pp. 1-2). However, there is little research that investigates employee AI literacy, especially empirical studies (Long and Magerko 2020). Therefore, we aim to develop a taxonomy that classifies relevant dimensions and characteristics of AI literacy and explore the concept conceptually and empirically. Our research question is: *What constitutes AI literacy?*

We develop a taxonomy of AI literacy using the methodological framework suggested by Nickerson et al. (2013). Thereby, we aim to organize the dimensions and characteristics comprising AI literacy in a

conceptual framework. We adopt an iterative taxonomy development process comprising repeated rounds of conceptual-to-empirical iterations. We report the results of the first iteration's conceptual phase and synthesize previous knowledge on AI literacy and related literacy concepts to derive deductively a conceptualization of what constitutes AI literacy. We thus classify initial AI literacy dimensions and their characteristics. Further, we present the results of the first iteration's empirical phase and revise our initial AI literacy taxonomy empirically based on interviews. This phase resembles a pilot study for the interviews that will be conducted during the second iteration of our conceptual-to-empirical taxonomy building process. Finally, we report first results on the conceptual phase of the second iteration in terms of an updated literature review. We offer two contributions to Information System (IS) research: we extend previous conceptualizations of AI literacy, both theoretically and empirically; and we begin to establish a taxonomy for AI literacy that can enhance theory.

In the section that follows, we outline the theoretical background. We then describe the taxonomy development process, with its underlying methods, as our main research approach. In the third section, we report on the first iteration's conceptual phase combined with insights from the second conceptual phase, followed by the results of the first empirical phase. The paper concludes with an outlook.

Theoretical Background

The idea of *literacy* has been studied intensely in education and technology (McLean 2013). Awareness of the benefits and disadvantages of technology and the ability to develop a critical understanding in a technologically mediated world is known as *technological literacy* (Dakers 2006, p. 2). Technological literacy can be understood as an overarching concept comprising forms of literacy, such as AI literacy. As AI is a specific form of technology, it is subject to *technological literacy*. Further, AI is related to other literacies frequently discussed in IS research (Long and Magerko 2020), such as *data literacy*, that is, the ability to analyze and work with data as part of a broader process of inquiry into the world (D'Ignazio 2017, p. 7), and *digital literacy*, which describes the ability to use information and communication technologies (ICT) effectively (Koltay 2011).

Computer literacy and *computational literacy* are additional literacies pertaining to AI. Computer literacy concerns the skills required to use a computer and computational devices in general, whereas computational literacy is a broader concept characterized by possessing skills that enable individuals to understand software and programs to work with those devices in detail (DiSessa 2001, pp. 4-5). Like AI literacy, they belong to the family of technological literacies. As AI is increasingly embedded in computational devices, AI-based technologies are also subject to computer literacy (Leander and Burriess 2020). However, AI likely requires more than just computational literacy-related skills (Leander and Burriess 2020).

While AI has been discussed in several literacy studies, the unique characteristics of human-AI interaction warrant considering AI as a specific concept (Leander and Burriess 2020). To make full use of AI, creative thinking and judgment are highly relevant (Plastino and Purdy 2018). In addition to technical skills, attitudes and ethics must also be considered (Finster and Robra-Bissantz 2020). Long and Magerko (2020) introduced *AI literacy* as a broader concept, based on existing research, of the competencies required as AI transforms how we communicate and work. They define AI literacy as “a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace” (Long and Magerko 2020, p. 2). Following this definition, AI literacy is considered on the individual level, which fits our aim of establishing an understanding of employee AI literacy.

Methods

Our goal is to improve the conceptual foundation of AI literacy. In pursuit of this goal, we develop a taxonomy of AI literacy based on the methodology suggested by Nickerson et al. (2013). Taxonomy development is an appropriate approach to achieve our goal, as it allows us to differentiate clearly and relate literacies pertaining to AI literacy in a comprehensive conceptual framework. Following Nickerson et al.'s guidelines, we adopted a conceptual-to-empirical approach to taxonomy development comprising a conceptual and an empirical phase. In this short paper, we present results from both phases of the first conceptual-to-empirical iteration, and first insights into the second iteration. In the first iteration's conceptual phase, we performed a systematic literature review to deduce an initial taxonomy of AI literacy based on related literature. Based on the results, we conducted semi-structured interviews during the empirical phase, which serves as a pilot study for the empirical phase in subsequent iterations. The second iteration's conceptual phase resembles an update of the literature review conducted for the first iteration. In the following section, we provide details on the methods used in each phase.

Conceptual Phase: Systematic Literature Review

In the conceptual phase, we established an initial classification of the AI literacy dimensions and their underlying characteristics. The overview given through existing theory will be an important source for developing our taxonomy. In doing so, we conceptualize AI literacy deductively. First, we explicate initial AI literacy dimensions based on the literacy literature in a technological context, such as digital literacy. To classify the underlying characteristics of each AI literacy dimension, we conducted a literature review following the guidelines of Templier and Paré (2018). We extended the identified dimensions and characteristics by including the literature of related literacy areas, such as computer/computational literacy and data literacy. Thereby, we aim to establish a robust and concise taxonomy that examines all relevant objects. The result is a deductively derived initial AI literacy taxonomy that serves as a foundation for the qualitative study in the first iteration's empirical phase.

As Nickerson et al. (2013) state, a taxonomy should be comprehensive and provide all current objects within the domain under consideration. Therefore, we chose a literature review to classify all known objects in the context of AI literacy in the first conceptual phase. We conducted a theoretical literature review, as it is appropriate for developing novel conceptualization (Paré et al. 2015). Our theoretical review focuses on the identification, description, and transformation of relevant AI literacy objects into the structure of taxonomy dimensions and their characteristics.

We searched for relevant literature by querying the title, abstract, and keywords – if available – for the keywords “Artificial Intelligence literacy,” “AI literacy,” and “AI literate” in the following databases: AIS Electronic Library, Business Source Ultimatum via EBSCOhost, Web of Science, ACM Digital Library, and JSTOR. Due to the novelty of the topic, we did not restrict our search to peer-reviewed literature, quality criteria (such as outlet-rankings), or certain time ranges. Our first search includes all articles published before March 2021. It yielded 42 unique hits across all databases and 27 hits after eliminating incomplete (5x) articles and grey literature (10x). We applied further criteria of relevance by eliminating non-AI-related articles (1x) and articles that did not make their own contributions to the conceptualization of AI literacy (22x including 1 design-oriented article). Some 4 articles remained for the taxonomy development. We determined that no taxonomy of AI literacy has been proposed to date. Almost all identified relevant articles refer to Long and Magerko (2020) when defining AI literacy (e. g., Long et al. 2021; Druga and Ko 2021). As already explained above, this is why we used their definition as our meta-inference in the taxonomy development process of Nickerson et al. (2013). As part of the second conceptual-to-empirical round, the review was repeated to identify articles published between March and August 2021. In this round we identified 5 additional relevant articles.

Empirical Phase: Semi-Structured Interviews

In the first iteration's empirical phase, we investigated AI literacy empirically through semi-structured interviews, which led us to revise our initial taxonomy. Following Nickerson et al. (2013) our goal was to find similarities and differences in the empirical data that would help us enhance or reframe the classification. The exploratory focus of the qualitative interviews (Creswell and Poth 2016, pp. 73-75) allowed us to investigate AI literacy from the employee perspective. The iterative changes we applied in the first conceptual-to-empirical iteration demonstrate the extendibility of our taxonomy by including further dimensions and characteristics as one of Nickerson et al. (2013) criteria for qualitative taxonomy development.

We conducted our in-depth interviews in April 2021 in a semi-structured way, following the guidelines for qualitative interviewing of Myers and Newman (2007). We used the conceptualization from the first conceptual phase to structure the interviews. We focused on open-ended questions so interviewees could express themselves freely. The interviews had four sections: the first section included general questions about the firm and the interviewee's position, responsibilities, experience with AI, and technological background; questions in the three other sections corresponded to the three dimensions of AI literacy discussed above. The interview questionnaire was revised after conducting the pilot study.

The interviews in our pilot study were conducted in online videoconferences and averaged 50 minutes in length. Participants were acquired via social media postings and organizational contacts. This selection approach should ensure great heterogeneity of the participants (Lee et al. 1999). We conducted five interviews in total with employees of five different companies in the finance, automotive, manufacturing, information and communication, and engineering industries. We chose employees who were rather tech-savvy and who we thought might be best for evaluating our interview design.

After transcribing our interviews, we used MAXQDA for coding. For qualitative content analysis, we used a deductive and inductive coding process. In the first step, we followed the steps for deductive category application of Mayring (2014, pp. 11-15) and worked with our previously formulated theory

components by bringing them into a coding agenda for main and sub-categories of codes. In the second step, we added further inductive categories by combining old categories or adding new ones to capture additional aspects of employee AI literacy based on the interviews (Mayring 2014, pp. 10-11).

In advancing with our work (but not a part of this short paper), we use literature to explain the nature of each empirical cluster, continuing to employ an empirical-to-conceptual approach. In doing so, we follow a cycle until all objects are defined precisely with ending conditions and, therefore, the taxonomy is completed (Nickerson et al. 2013).

Toward a Taxonomy of AI Literacy: Conceptual Phase

In this section, we focus on the conceptual phases from the first and second iteration of the taxonomy development process of Nickerson et al. (2013) by identifying dimensions of AI literacy and its underlying characteristics. We explain AI literacy dimensions based on the related literature and differentiate it from technological and digital literacy. We aim to develop unique dimensions as one of our ending conditions. Further, we derive characteristics for each AI literacy dimension based on the literature review, following Templier and Paré (2018). We enhanced the characteristics by integrating theory from related literacies, such as computational and data literacy, given that the literature on AI literacy is at an early stage (Long and Magerko 2020). The result is a deductively derived conceptualization of AI literacy that serves as the foundation for our qualitative study in the second iteration and further taxonomy development. In the following, we present the combined results both conceptual phases, as the second phase did not result in changes to the taxonomy.

AI literacy dimensions

Among the literacies discussed in the context of technology in general, three are particularly relevant to AI literacy: functional/technical literacy, critical/cognitive literacy, and sociocultural/socioemotional literacy (McLean 2013; Ng 2012). Existing work on these literacies does not translate seamlessly to AI literacy. *Functional/technical literacy* concerns the skills individuals need to work with technologies and use them as powerful tools (Luke 1994, 2, 35, 39; McLean 2013; Ng 2012). AI-based technologies are evolving rapidly (Finster and Robra-Bissantz 2020), as are functional AI literacy skills. *Critical/cognitive literacy* is constituted through knowledge and experiences that enable individuals to interpret critically their experiences with technology (Luke 1994, 3, 32-33, 37; McLean 2013; Ng 2012). As individuals often have those experiences with AI outside of the workplace (Leander and Burriss 2020), we must consider both the work and non-work environments when exploring what constitutes critical AI literacy. *Sociocultural/socioemotional literacy* refers to an individual's general attitude about technology based on the embeddedness of technology in that individual's cultural environment and underlying social practices (Luke 1994, 2,6; McLean 2013; Ng 2012). Hence, values and beliefs from the environment underpin an individual's inherent understanding of technology (Luke 1994, p. 16). These three types serve as our AI literacy taxonomy dimensions, having distinguishing characteristics differentiating them (Nickerson et al. 2013).

While the AI literacy definition of Long and Magerko (2020) encompasses functional and critical literacy to some extent, sociocultural literacy is not represented well. Both "use AI as a tool" and "collaborate effectively with AI" in the definition relate to functional AI literacy, whereas "critically evaluate AI technologies" concerns critical AI literacy. Any allusion to sociocultural AI literacy in the definition is limited to their mention of different environments of use. Quite similar to Long and Magerko (2020), Dai et al. (2020, p. 3) define AI literacy as the "ability to access and use AI-related knowledge and skills." Due to the educational context of their work, their primary focus is on functional literacy skills. Hermann (2021, p. 13) added critical literacy abilities to functional AI literacy skills by including the individual's basic understanding of "AI's susceptibility to biases and selectivity" and "AI's potential impact." Chai et al. (2020, p. 6) constructed AI literacy with functional and critical AI literacy skills.

The debate over whether technical competencies make individuals AI literate is also based on K–12 competencies. For instance, Chai et al. (2021) operationalize AI literacy through the elementary school's AI curriculum and, therefore, provide only the functional dimension of AI literacy. Long et al. (2021) and Sambasivan et al. (2021) refer to functional AI literacy using AI competencies identified in the literature as well as AI skills that are relevant for the K–12 audience. In comparison, Druga and Ko (2021) approached children's AI literacy through the lens of sense-making as an individual and social process of individuals encountering unfamiliar situations. They offer some sociocultural-related thoughts regarding AI literacy. Williams et al. (2021) discuss sociocultural AI literacy by using the theory of minds based on the ability to reflect AI from another standpoint. The sociocultural dimension of AI literacy plays a significant role, especially as AI requires communication and creative thinking rather than technical skills alone (Plastino and Purdy 2018).

In the following section, we combine the work of Long and Magerko (2020), as most of the literature identified in our review refers to their conceptualization with the elements of critical, functional, and sociocultural literacy of McLean (2013). The result is an initial conceptualization of AI literacy dimensions using the conceptual-to-empirical approach of Nickerson's taxonomy development (2013).

Characteristics of the AI literacy dimensions

In this section, we extend each AI literacy dimension by spelling out its characteristics. We first introduce the core AI competencies in the conceptualization of Long and Magerko (2020), who already summarized the theoretical discourse on AI literacy to date, and we then enhance this with insights from the relevant literature identified in our review. We add potential measures from other literacy areas that fit well in the scope of AI literacy. This represents a continuation of the conceptual-to-empirical approach of the taxonomy development of Nickerson et al. (2013) used in our first iteration.

Functional AI literacy

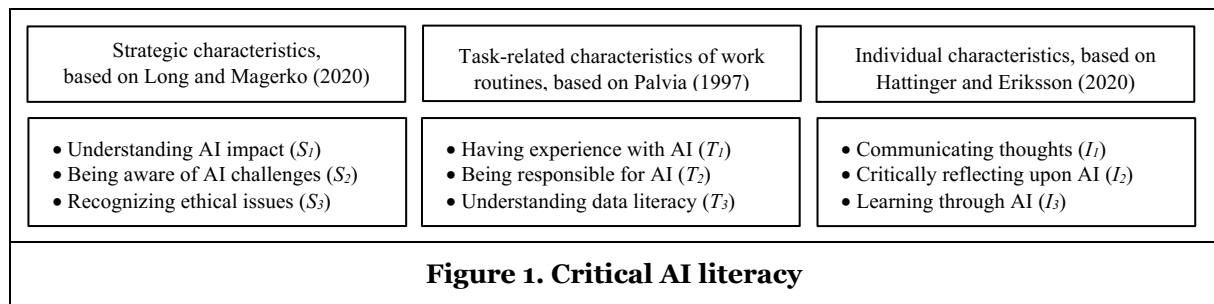
Long and Magerko (2020) spell out several AI literacy competencies based on their review of the literature. We relate these competencies to functional AI literacy, as they are necessary to work with AI-based systems or use AI-based tools effectively. They include knowing what AI is (K_1), such as recognizing AI and a variety of AI-based systems; and distinguishing between human and machine intelligence. They also include knowing how AI works (K_2) by understanding knowledge representation and how systems make AI-based decisions. The authors reference robotics in their conceptualization of AI literacy, as they review work that spans from robotics into AI more broadly. Functional AI literacy should include knowing that some actions of AI-based systems may be carried out physically by robots and that robots are programmed (K_3). Those characteristics of functional AI literacy are similar to the AI literacy construct of Chai et al. (2020), William's theory of mind AI skills (2021), and the AI concepts of Sambasivan et al. (2021). In addition, we consider the components of computer literacy skills articulated by Palvia (1997) to be part of AI literacy. These include self-taught AI skills (E_1), formal education (E_2), and AI training (E_3), which we assign to functional AI literacy because they concern the skills individuals need to work with AI.

Critical AI literacy

Understanding what AI can do by identifying AI challenges, using AI appropriately, and imagining future AI applications are aspects of the Long and Magerko (2020) set of core competencies of AI literacy that we assign to critical AI literacy, as they concern evaluating the consequences of interacting with AI. In addition, there are machine learning-related competencies that fit with the concept of critical AI literacy, such as understanding the challenges of machine learning, recognizing the human role in programming, understanding basic data literacy concepts, recognizing that computers learn from data, and interpreting data critically. As mentioned above, data literacy involves using data as part of a broader process (D'Ignazio 2017). The definition of data literacy overlaps with AI literacy (Long and Magerko 2020). The authors also include as a core competency of AI literacy the ability to identify and describe different perspectives of key ethical issues surrounding AI; we assign it to critical AI literacy. Hermann 2021 and Sambasivan et al. (2021) also demonstrate the relevance of ethical abilities in the context of AI literacy. Further, Druga and Ko (2021) mention the relevance of social judgment and agency, such as the manner in which the agent is controlled.

Of the computer literacy skills Palvia (1997) identifies, we assign experience and responsibility to critical AI literacy, as these competencies enable employees to interpret AI's impact. Experience can be further specified using the five components of Seyal et al. (2001): experience in using different AI-based systems; being able to build AI models; experience in programming or establishing AI; participating in non-technical AI design studies; and participating in technical AI studies. As reflection is a critical factor for learning mechanisms (Hattinger and Eriksson 2020), individuals also require adaptive learning abilities through self-evaluation to interact with technology. Learning through AI has great potential for personal (Hermann 2021) and interest development (Long et al. 2021), and so creativity is important to explore AI activities (Long et al. 2021). This also points to data literacy as a concept with similarities to AI literacy (Long and Magerko 2020), because it creates the ability to interpret a piece of information for engaging in a broader process of inquiry into the world (D'Ignazio 2017, p. 7).

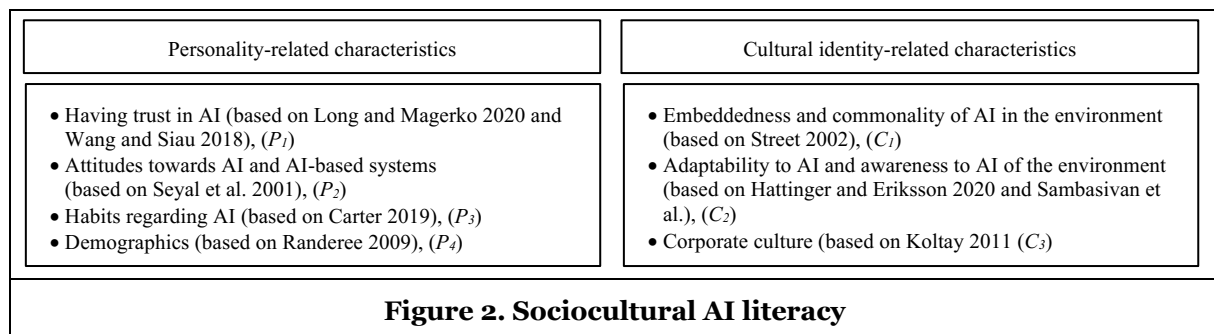
In summary, we have strategic characteristics (such as thinking about ethics) as well as specific task-related characteristics of AI literacy in the daily work routine (such as being careful when entering or interpreting data). In addition, we account for individual-related elements of critical AI literacy, which are not tied to specific tasks; these are individual competencies (such as critically reflecting upon interactions with and use of AI technologies). Figure 1 summarizes all three characteristics.



Sociocultural AI literacy

While the conceptualization of Long and Magerko (2020) does not explicitly spell out sociocultural competencies, the authors do raise related characteristics. For instance, they argue that knowledge is important to trust in AI. Wang and Siau (2018) identified factors that affect trust building in the use of chatbots, such as predictability and transparency. An individual's trust can pertain to sociocultural AI literacy. Seyal et al. (2001) introduced several attitudes in the context of computer literacy that can be related to sociocultural AI literacy. Going further, Carter (2019) determined that habits also affect how individuals work with technologies. The author mentions that it can be difficult for professionals who are becoming habituated to searching through and manually looking at data to let a machine decide, whereas the non-professional end user is more likely to embrace the automatic refining abilities when searching. Hence, habits contribute to sociocultural literacy. Younger generations may be more suited to collaborating with AI, given that they are more tech-savvy and tend to encounter AI in school (Randeree 2009). Thus, demographics may be also relevant for sociocultural AI literacy.

The elements mentioned thus far relate to characteristics grounded in an individual's personality and demographics. There are also elements that are more related to cultural identity, such as the embeddedness of AI in the cultural environment, that is, how common it is to use AI-based technologies at work and in private environments. Street (2002, p. 79) cautions that the social context can reinforce a certain ideology and restrict an individual from critically evaluating a situation, such as hearing stories about experiences with AI from friends and colleagues (Druga and Ko 2021). This shows the importance of sociocultural AI literacy. Also, the adaptability and awareness of the employee's environment are relevant for learning how to adapt to AI (Hattinger and Eriksson 2020; Sambasivan et al. 2021). Furthermore, Koltay (2011) shows a relationship between culture and literacy. To sum up:



Toward a Taxonomy of AI literacy: Empirical Phase

In this section, we discuss our empirical phase of the first iteration of the taxonomy development process of Nickerson et al. (2013). We used the conceptualization from the first conceptual phase to guide the execution of the exploratory qualitative interviews (Creswell and Poth 2016, pp. 73-75). This is aimed at enhancing our AI literacy taxonomy empirically by exploring possible differences between employees with low and high levels of literacy. In this way, we extend and revise our initial taxonomy classification from the first conceptual phase. Given our emphasis on understanding AI literacy within the context of augmenting an employee's work, we employed an exploratory study (Creswell and Poth 2016, pp. 73-75) that allowed us to explore AI literacy from the employee's perspective (Lee et al. 1999). Table 1 provides insights into the pilot coding.

Dimension	Functional AI literacy						Critical AI literacy									Sociocultural AI literacy								Novel character- istics (inductive)
Characteristic (main codes)	Content			Path- way			Strateg- ic			Task			Individ- ual			Personality				Cultural identity				
Sub-codes (deductive)	K ₁	K ₂	K ₃	E ₁	E ₂	E ₃	S ₁	S ₂	S ₃	T ₁	T ₂	T ₃	I ₁	I ₂	I ₃	P ₁	P ₂	P ₃	P ₄	C ₁	C ₂	C ₃		
Interview 1	x		x		x	x		x	x	x	x	x		x		x	x		x	x	x	x		
Interview 2	x	x		x		x		x		x	x		x		x	x	x		x	x		x		
Interview 3	x		x	x	x		x	x	x	x			x	x		x		x	x	x	x	x		
Interview 4	x		x		x	x		x	x	x			x	x	x	x	x	x	x	x	x	x		
Interview 5	x	x		x	x	x	x	x	x	x	x		x	x	x	x	x		x	x		x		
Table 1. Coding insights																								

Preliminary Results and Discussion

In testing our questionnaire through the pilot, we determined that the questions were largely appropriate given that we were able to relate the answers to our previous conceptualization. The open-ended questions allowed us to go further in our interpretation. Here we derive some limited preliminary results.

First, interviewee definitions of AI were similar to our underlying definition (Rai et al. 2019), that is, as an augmentative system that supports its users by performing cognitive functions, such as making decisions based on experiences represented by data. Most interviewees stated that they do not use the term “AI” often, but rather refer to analytics or specific forms of AI processes, as there is no single “true” AI. Our interviewees mentioned that employees in other non-technical units in their organizations do not have a good understanding of AI and conceive of it in a physical sense as a way of making it more tangible, such as seeing AI as a robot that takes over the functions of a bank teller. Furthermore, those same non-technical employees think about AI as a buzzword for automated intelligence in form of a robot, while not recognizing the AI-based systems already being used in daily life and work. The interviewees argued that their education and their professional experiences dealing with AI make them able to understand AI better than their colleagues who lack technical backgrounds. This could indicate that there is an educational characteristic of functional AI literacy. In addition, their answers indicate that it might be useful for organizations to establish an organizational understanding of what is meant by AI. They expressed the need for common definitions of AI in their companies that spell out a clear vision of how the firm and its employees will use AI-based tools; this would enhance the understanding of non-technical employees and provide clear guidance. This could indicate that there are cultural identity-related characteristics of sociocultural literacy that might influence functional AI literacy.

Second, all our interviewees referenced competencies that are relevant irrespective of the particular AI-based system being used. They mentioned hard skills, such as a knowledge of statistics that would enable interpreting data and seeing correlations, as well as soft skills. Being creative, willing to learn, and able to communicate are important soft skills given that working with AI requires “all-around” competencies beyond just specific programming skills. Furthermore, the interviewees mentioned that they often ask existing sources of knowledge, such as colleagues, or utilize open source software, rather than trying to learn everything in detail on their own—which they said would take too long. This combination of different knowledge sources could demonstrate how different individuals’ AI literacy correlates in an organizational context. Furthermore, some interviewees described themselves as curious or open to new things and generally related this characteristic to their generation (they were around 30 years old), in slight contrast to the older interviewees, who instead highlighted interpersonal characteristics such as empathy regarding the human component of AI use, as well as morality concerns. Hence, demographics may matter in sociocultural AI literacy. Interviewees also differentiated between non-work and work-related use of AI-based systems. They reported that using the same AI tool at work requires more background knowledge than using it in a private context (e.g., due to responsibilities and data security concerns at work). This could be a task-related characteristic of critical literacy. Further, interviewees stated that individuals should be able to use the hardware and change configurations. That is a new usage-related point we consider part of functional AI literacy.

Our final insight is that our interviewees indicated that corporate culture influences critical AI literacy. Interviewees working in smaller companies with a rather homogenous workforce and a less-regulated culture tended to be less questioning of the impact of using AI, and evaluate the influence on a personal level, whereas those working in firms with a high security orientation and heterogeneous employee

groups often discussed the effects of using AI not only on themselves but also on others in their organizations. Hence, besides individual aspects of critical AI literacy, there could also be organizational aspects. Furthermore, sociocultural AI literacy might be interwoven with critical AI literacy.

To sum up, we identified several insights that strengthen and extend our conceptualization of the functional, critical, and sociocultural AI literacy dimensions. Moreover, we determined potential correlations between the AI literacy dimensions. Furthermore, there are indicators of how the AI literacy elements at the individual level could help constitute an AI-literate workforce at the organizational level. In addition, we observed that sociocultural AI literacy plays a huge role in constructing overall AI literacy.

Conclusion and Outlook

The growing augmentation at work through AI (Rai et al. 2019) requires understanding the factors that influence AI literacy. This goes beyond simply looking at competencies to include factors such as trust and culture, for instance. We extend the foundations of AI literacy by synthesizing, both conceptually and empirically, characteristics of functional, critical, and sociocultural literacy into an AI literacy taxonomy. We will continue our qualitative interviews with heterogeneous employee groups, particularly with non-technical workers, to explore what constitutes employee AI literacy and why there are differences among employees. AI literacy can facilitate the successful collaboration of employees and AI-based systems. As the pilot study determined, it may be important to clarify the organizational understanding of AI and provide guidance on usage. Understanding AI literacy also has implications for change and knowledge management practices. Finally, we should state that this short paper is based on a very limited number of interviews. Further, the theoretical framing of our conceptualization is limited by the fact that we synthesized only specific literature. As we continue our work, we will explore the interdependencies of the AI literacy dimensions to help inform which variables will have what sort of impact on building an AI-literate workforce.

References

- Carter, D. 2019. "New technologies and new data sources: The business information survey 2019," *Business Information Review* (36:3), pp. 95-105.
- Chai, C. S., Lin, P. Y., Jong, M. S. Y., Dai, Y., Chiu, T. K., and Qin, J. 2021. "Perceptions of and Behavioral Intentions towards Learning Artificial Intelligence in Primary School Students," *Educational Technology & Society* (24:3), pp. 89-101.
- Chai, C. S., Wang, X., and Xu, C. 2020. "An Extended Theory of Planned Behavior for the Modelling of Chinese Secondary School Students' Intention to Learn Artificial Intelligence," *Mathematics* (8:11), pp. 2-18.
- Creswell, J. W., and Poth, C. N. 2016. *Qualitative inquiry and research design: Choosing among five approaches*, Los Angeles, London, New Delhi: SAGE publications.
- Dai, Y., Chai, C.-S., Lin, P.-Y., Jong, M. S.-Y., Guo, Y., and Qin, J. 2020. "Promoting Students' Well-Being by Developing Their Readiness for the Artificial Intelligence Age," *Sustainability* (12:16), p. 6597.
- Dakers, J. R. 2006. "Introduction: Definition Technological Literacy," in *Defining Technological Literacy: Towards an Epistemological Framework*, J. R. Dakers (ed.), New York: Palgrave Macmillan.
- D'Ignazio, C. 2017. "Information Visualization," *Information Design Journal* (23:1), pp. 6-18.
- DiSessa, A. A. 2001. *Changing minds: Computers, learning, and literacy*, Cambridge: MIT Press.
- Druga, S., and Ko, A. J. 2021. "How do children's perceptions of machine intelligence change when training and coding smart programs?" in *Interaction Design and Children*, M. Roussou, S. Shahid, J. A. Fails and M. Landoni (eds.), New York: ACM, pp. 49-61.
- Finster, R., and Robra-Bissantz, S. 2020. "Not just Skills and Knowledge: Fostering Competencies in Information Systems Education," *Proceedings of Americas Conference on Information System*, p. 2.
- Hattinger, M., and Eriksson, K. 2020. "Mind the Gap: A Collaborative Competence e-Learning Model between University and Industry," *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pp. 79-88.
- Hermann, E. 2021. "Artificial intelligence and mass personalization of communication content: An ethical and literacy perspective," *New Media & Society*, pp. 1-20.
- Koltay, T. 2011. "The media and the literacies: Media literacy, information literacy, digital literacy," *Media, Culture & Society* (33:2), pp. 211-221.
- Leander, K. M., and Burris, S. K. 2020. "Critical literacy for a posthuman world: When people read, and become, with machines," *British Journal of Educational Technology* (51:4), pp. 1262-1276.

- Lee, T. W., Mitchell, T. R., and Sablinski, C. J. 1999. "Qualitative Research in Organizational and Vocational Psychology, 1979-1999," *Journal of Vocational Behavior* (55:2), pp. 161-187.
- Long, D., and Magerko, B. 2020. "What is AI Literacy? Competencies and Design Considerations," *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-16.
- Long, D., Padiyath, A., Teachey, A., and Magerko, B. 2021. "The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences," *Creativity and Cognition*, pp. 1-10.
- Luke, A. 1994. *The social construction of literacy in the primary school*, Macmillan Education AU.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., and Söllner, M. 2019. "AI-Based Digital Assistants," *Business & Information Systems Engineering* (61:4), pp. 535-544.
- Mayring, P. 2014. *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*, Klagenfurt.
- McLean, K. 2013. "Literacy and Technology in the Early Years of Education: Looking to the Familiar to Inform Educator Practice," *Australasian Journal of Early Childhood* (38:4), pp. 30-41.
- Myers, M. D., and Newman, M. 2007. "The qualitative interview in IS research: Examining the craft," *Information and Organization* (17:1), pp. 2-26.
- Ng, W. 2012. "Can we teach digital natives digital literacy?" *Computers & Education* (59:3), pp. 1065-1078.
- Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A method for taxonomy development and its application in information systems," *European Journal of Information Systems* (22:3), pp. 336-359.
- Palvia, P. C. 1997. "Developing a Model of the Global and Strategic Impact of Information Technology," *Information and Management* (32:5), pp. 229-247.
- Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. 2015. "Synthesizing information systems knowledge: A typology of literature reviews," *Information & Management* (52:2), pp. 183-199.
- Plastino, E., and Purdy, M. 2018. "Game changing value from Artificial Intelligence: eight strategies," *Strategy & Leadership* (46:1), pp. 16-22.
- Rai, A., Constantinides, P., and Sarker, S. 2019. "Editor's comments: next-generation digital platforms: Toward human-AI hybrids," *MIS Quarterly* (43:1), pp. iii-x.
- Randeree, E. 2009. "Investigating literacy as a predictor of adoption and usage," *Proceedings of the Southern Association for Information Systems Conference*, pp. 152-157.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work: Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-15.
- Seyal, A., Rahman, N., and Mohamad, H. 2001. "Organizational Characteristics, Computer Literacy and Attitudes of CEOs' toward the Use of Information Technology: A Study of Bruneian Small & Medium Business Enterprises," *PACIS 2001 Proceedings*, p. 57.
- Street, B. V. 2002. *Literacy and development: Ethnographic perspectives*, London: Routledge.
- Templier, M., and Paré, G. 2018. "Transparency in literature reviews: an assessment of reporting practices across review types and genres in top IS journals," *European Journal of Information Systems* (27:5), pp. 503-550.
- Wang, W., and Siau, K. 2018. "Living with Artificial Intelligence: Developing a Theory on Trust in Health Chatbots," *Proceedings of the Sixteenth Annual Pre-Workshop on the International Conference on Information Systems*, pp. 1-5.
- Williams, R., Park, H. W., and Breazeal, C. 2021. "A is for Artificial Intelligence," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-11.

Paper VI

Heyder, T; Passlack, N. & Posegga, O. (2023) *Ethical Management of Human-AI Interaction: A Theory Development Review*. The Journal of Strategic Information Systems (JSIS), 32(3), 101772, <https://doi.org/10.1016/j.jsis.2023.101772>.

Ethical Management of Human-AI Interaction: Theory Development Review

Completed Manuscript – Accepted Version

1. Introduction

The most recent generations of technologies based on Artificial Intelligence (AI), characterized by the ability “to perform cognitive functions that we associate with human minds” (Rai et al., 2019, p. iii), change the nature of the symbiosis between humans and technology: AI increases “human task performance with higher extents of interactivity and intelligence” than previous technologies (Maedche et al., 2019, p. 536). AI differs from previous technologies in its ability to act (semi-)autonomously (Maedche et al., 2019; Rieder et al., 2020; Scherer, 2016). For example, recent developments in generative AI, such as the growing maturity of Large Language Models (LLMs), increase the relevance and acceptance of AI-based technologies in organizations (Dwivedi et al., 2023; Markus and Rowe, 2023). With their growing capabilities, maturity, and persuasiveness, such technologies have the potential to reshape traditional forms of work, giving rise to “human-machine hybrid work” (Mollick, 2022, p. 5). AI-based technologies are “no longer always subordinate to the human agent, [as they] can now assume responsibility for tasks” (Baird and Maruping, 2021, p. 315). When AI-based systems act (semi-) autonomously (Rieder et al., 2020), autonomy shifts from humans to technical systems. This influences the collaboration of humans and technology (Baird and Maruping, 2021), raising ethical challenges such as distributive justice concerns, discrimination and exclusion, and transparency issues through the perceived powerlessness of humans as well as AI opacity (e.g., Abdelaal, 2021, Boada et al., 2021, Dwivedi et al., 2023, Giermindl et al., 2022).

To avoid unintended negative consequences of the ethical challenges that result from this changing nature of human-AI interaction, we must rethink how we manage human-AI interaction in organizations (Rai et al., 2019). To promote positives outcomes such as well-being, organizations must consider ethics and morals when making decisions about the implementation and use of the latest technologies based on AI to realize their strategic goals (Berente et al., 2021; Dwivedi et al., 2023; Marabelli et al., 2021). Considering ethics in strategic management requires two perspectives. On the one hand, managing human-AI interaction needs to enforce

ethics by duty by establishing guidelines and developing principles that prescribe correct ethical behavior as an obligation for employees and AI (Alsheibani et al., 2020; Han et al., 2020; Rakowski et al., 2021). On the other hand, strategic management needs to be aware of the workforce's *virtue ethics* based on the inherent moral values that guide the intrinsic motivation of individuals to behave ethically when implementing, designing, using, and deciding upon AI (Flathmann et al., 2021). Hence, to foster ethical management of human-AI interaction, it is important to combine both ethical perspectives (Eitel-Porter, 2021; Siau and Wang, 2020).

Although recently there has been valuable research on ethics in the context of AI, there appear to be four gaps in research on the ethical management of human-AI interaction. First, the discourse lacks a general theory that explains the complexity of ethics and AI (Mirbabaie et al., 2022), that is, combining the technical discourse with ethical philosophy (Berente et al., 2021). Second, research has provided limited approaches to understanding contemporary ethical challenges at the crossroads of society, business, and technology (Islam and Greenwood, 2021). Third, the current discourse on ethics in the context of managing human-AI interaction is restricted mainly to the formulation of “principles, rules, guidelines, policies, and regulations for AI ... to exhibit ethical behavior” (Siau and Wang, 2020, p. 74), representing only one of the perspectives (*duty ethics*) required to integrate ethics in strategic management. Finally, Information Systems (IS) research and research in closely related fields often adopt a technical stance and gravitate toward studying and designing technologies, including AI methods and algorithms, as well as their characteristics, such as fairness and explainability. They often fail to consider explicitly human actors and how their behavior might affect human-AI interaction, which is also required to understand and resolve ethical issues in workplace management of human-AI interaction (Amershi et al., 2019; Berente et al., 2021; Te'eni et al., 2019). In addressing those four gaps, we strive for an interdisciplinary, systematic, and holistic approach that addresses the following research question:

How can we understand and explain ethical management of human-AI interaction?

In order to do justice to the complexity of the phenomenon under research and to ensure a human-centered perspective, there is a need to reflect upon the management of human-AI interaction in both directions: managing AI interacting with humans (incorporating the *duty ethics* perspective of integrating ethical principles in the design of AI-based technologies); and managing humans interacting with AI (incorporating the *duty ethics*

perspective of establishing ethical principles to guide their behavior when working with AI as well as the *virtue ethics* perspective aiming to align human values with AI). This leads to the two subquestions (SQs): *How can we understand and explain ethical management of AI interacting with humans?* (SQ1), and *How can we understand and explain ethical management of humans interacting with AI?* (SQ2).

To achieve our research goal, we conduct a *theoretical review*. This type of literature review aims at building an explanation by employing a systematic and structured approach to transforming constructs identified in previous work, spread across several research streams, into a conceptual framework that allows for examining its components and interdependencies (Paré et al., 2015). In terms of theory development, we adopt the *extend strategy* of “Next-Generation Theorizing” of Burton-Jones et al. (2021, p. 303) by using well-established theories to explain human-AI interaction as a “new phenomenon[on] emerging in a changing world” (Burton-Jones et al., 2021, p. 303). To do so, our work builds on the conceptualization by Leonardi (2012), which summarizes the discourse on sociomateriality in a conceptualization of sociotechnical systems. The sociotechnical perspective allows us to reflect upon the current progress of AI-based technologies, as such technologies interfere with interpersonal relations and, thus, represent “sociotechnical systems par excellence” (Greene et al., forthcoming, p. 2). Sociotechnical system theory represents how interdependencies between humans, technologies, and other objects (such as guidelines and management practices) being arranged in organizations lead *inter alia* to ethical or non-ethical patterns of actions (Cecez-Kecmanovic et al., 2014; Contractor et al., 2011). We extend Leonardi’s (2012) conceptualization through state-of-the-art literature on managing human-AI interaction in organizations that we identified in our systematic literature review. To provide a more nuanced analysis of this literature, we use the notions of sociomateriality (*inseparability, relationality, performativity, materiality, and practices*) as they “individually and collectively, offer a distinctive and coherent account of the relationships between” all parts (human, technologies, and other objects) being arranged in organizations (Jones, 2014, p. 895). These notions enable us to improve our understanding of ethical interdependencies arising through the “entanglement” of humans and technologies in organizations influenced by management practices and guidelines (Cecez-Kecmanovic et al., 2014). We thus go beyond synthesizing the current literature, aiming to deliver a comprehensive *theoretical review* to increase knowledge on the ethical management of human-AI interaction involving both ethical perspectives (*duty* and *virtue*).

We expect to contribute to research in three areas. First, we extend knowledge on the ethical management of human-AI interaction systematically, not only by consolidating existing research insights but also by explaining as sociotechnical systems the ethical interdependencies between humans, technologies, and other objects in organizations. Second, we respond to the call of Berente et al. (2021) by applying the sociotechnical perspective and showing that it is not only a product of theorizing (outcome) but also can guide ethical considerations in managing human-AI interaction. We agree with Greene et al. (forthcoming), that the IS field is responsible for and “uniquely qualified to evaluate the scientific, social, and technical aspects” of the newest technologies being based on AI and using unsupervised learning techniques (p. 2). Thus, we provide a systematic approach for analyzing the ethical management of human-AI interaction through the lens of sociomateriality. Third, we combine the technical discourse with ethical philosophy to account for both the *duty ethics* and *virtue ethics* perspectives, and thus, following the calls of Berente et al. (2021) and Mirbabaie et al. (2022), for a general theory that accounts for the complexity of ethics and AI. Thereby, we combine older and new theories, not only from IS research but also from interdisciplinary domains, thus showing how “Next-Generation Theorizing” may be conducted in the context of a systematic *theoretical review* (Paré et al., 2015; Templier and Paré, 2018). Furthermore, we heed the calls of Vial (2019) and Alsheibani et al. (2020) to enhance our understanding of ethics with respect to the multilevel impact of digital transformation as a way to provide recommendations for the strategic management of AI that will contribute to enhancing the long-term performance of organizations.

In section 2, we outline sociomateriality and the sociotechnical system theory as theoretical lenses. We then describe, in section 3, our methodology and its two stages (*conceptual stage* and *review stage*). Section 4 is an elaboration of our conceptual framework as a result of the first research stage, the *conceptual stage*. Section 5 presents the results from the second research stage, the *review stage*. This is followed by a discussion of our results in section 6. The paper concludes in section 7 with a summary, limitations, and an outlook.

2. Theoretical Lens of Sociomateriality and Sociotechnical System Theory

As spelled out in the Introduction, we ground our work in that of Leonardi (2012), who applies the discourse on sociomateriality in a conceptualization of sociotechnical systems. Sociomateriality melds technology and

social phenomena to better describe how technology (as the material entity) is created and made observable through a variety of social phenomena (e.g., making decisions or formulating strategies as the social entity) that occur when the social and material entities are entangled in everyday practices. In contrast, sociotechnical systems describe the object of studying sociomateriality, such as organizations that are made up of the social (e.g., communication networks or status) and the material (e.g., technologies) (Leonardi, 2012). Sociomateriality draws on Actor-Network Theory (ANT), which can help further describe the power of the entanglement between the social and material entity. ANT considers how “network dynamics could ... shape what people thought about a new technology, as well as whether and how they would use it” (Contractor et al., 2011, p. 683); using the lens of sociomateriality that builds upon ANT is appropriate for identifying interdependencies between humans (incorporating *virtue ethics*), technologies, and other non-human objects, such as management practices (incorporating *duty ethics*) based on how they are entangled and arranged in organizations and how those interdependencies lead to different forms of associations and power (Baron and Gomez, 2016; Contractor et al., 2011). It helps explain how those interdependencies affect actions, such as ethical or non-ethical behavior, that emerge in the collection of human and non-human objects (Cecez-Kecmanovic et al., 2014; Contractor et al., 2011; Latour, 2005). Hence, we use Leonardi’s (2012) conceptualization to derive a theoretical lens grounded in sociomateriality, sociotechnical systems theory, and ANT.

Following Leonardi’s (2012) conceptualization, an organization as a sociotechnical system comprises the *social subsystem* and *sociomaterial practices* (sometimes referred to as the *technical subsystem*). *Sociomaterial practices* describe the entanglement of humans and technology, conceptualized as the social entity and the material entity (Orlikowski, 2007; Orlikowski and Scott, 2008). While humans decide how to respond to technology based on their (human) agency (Cooren, 2004, p. 377; Leonardi, 2011), which also is influenced by their ethical ideology and being (*virtue ethics*), the material agency can be described as “the way objects act when humans provoke it” (Leonardi, 2012, p. 37); this also involves their ethical actions as *duty ethics*. Further, *sociomaterial practices* shape and are shaped by “institutionalized ideas about how people could and should relate to one another” (Leonardi, 2012, p. 40), which can be described as the *social subsystem* (Leonardi, 2012). In the *social subsystem*, there can be several human objects (e.g., a manager’s way of leading or the workforce’s

culture, incorporating *virtue ethics*) and non-human objects (e.g., guidelines, practices, strategies, or other technological objects, incorporating *duty ethics*) that affect the ethical management of human-AI interaction. Thus, with a collection of human and non-human objects as part of the *social subsystem* and the social entity and material entity as part of *sociomaterial practices*, the lens of sociomateriality enables us to improve our understanding of their “entanglement” (Cecez-Kecmanovic et al., 2014, pp. 825–826).

In the following, we stick to the terms of Leonardi (2012) by using “social entity” when considering the human interacting with AI and “material entity” when talking about AI interacting with the human in *sociomaterial practices*. In comparison, when talking about the *social subsystem*, we refer to human objects (e.g., a manager’s way of leading or the workforce’s culture) and non-human objects (e.g., guidelines, practices, strategies, or other technological objects).

To explain and outline differences in actions and behavior resulting from the interdependencies in an organization, we use the five notions of sociomateriality first introduced in the work of Orlikowski (2010, 2007) and Orlikowski and Scott (2008) and subsequently discussed by Jones (2014), which account for the complexity of human-AI interaction in the sociotechnical system theory:

- i) *Inseparability* describes how the social and the material entities are interwoven;
- ii) *Relationality* describes how the entanglement of the social and material entity affects both entities (social and material entity) as an outcome;
- iii) *Performativity* describes how the social entity creates its reality;
- iv) *Materiality* describes how the material entity is arranged based on an artifact’s physical or non-physical objects; and
- v) *Practices* describe how the entanglement is embedded in the *social subsystem*.

The notions of sociomateriality can contribute to our understanding of how ethical interdependencies arise between the *social subsystem* and *sociomaterial practices*. For instance, managers can act as role models, advising and guiding employees to make more ethical decisions when collaborating with AI acting as role models (Alsheibani et al., 2020; Eitel-Porter, 2021). The notion *performativity* enables us to explain how managers’ *virtue ethics* (belonging to the *social subsystem*) are affected by how those managers are embedded

in organizations, as well as how their ethical behavior (based on their *virtue ethics* on a collective level) might influence how employees as social entities change their own reality on (*virtue*) ethics in human-AI interaction, leading to ethical or non-ethical behavior. Hence, we adapt Leonardi's (2012, p. 43) conceptualization of sociotechnical systems by applying the five notions of sociomateriality (*materiality, inseparability, relationality, performativity, and practices*) to analyze the literature from our review.

Given that the interdependencies between the arrangement of all entities (material and social) collected in the sociotechnical system affect their actions and behavior (Contractor et al., 2011; Latour, 2005), we must further be aware of the strength of those interdependencies – mainly as AI as the material entity increasingly acts autonomously (Baird and Maruping, 2021; Rieder et al., 2020), changing the nature of human-AI interaction in terms of *sociomaterial practices* (Cecez-Kecmanovic et al., 2014; Maedche et al., 2019). To account for this changing nature of human-AI interaction, we follow the arguments of Jones (2014) and differentiate between *weak* and *strong* notions of sociomateriality.

In general, *weak* sociomateriality assumes a mutual interdependency between the social entity (humans) and the material entity (AI), meaning that each entity, with preexisting and independent attributes, forms, and capabilities, has its own properties that do not change (Cecez-Kecmanovic et al., 2014). This view is also represented in the substantialist ontology as one possible ontology of sociomateriality. It is supported by Faulkner and Runde (2011), Leonardi (2012), and Mutch (2013), who argue that the social and material entities exist as “separate and self-contained entities that interact and affect each other” (Cecez-Kecmanovic et al., 2014, p. 809). Using our aforementioned example of managers as human objects guiding ethical human-AI interaction in organizations from the *social subsystem*, a *weak* form of the notion *performativity* is present when a manager's behavior may lead to changes in the ethical behavior of an employee but may not affect their ethical ideology (*virtue ethics*). Hence, the individual's *virtue ethics* has independent attributes that neither change through the external influence of managers from the *social subsystem*, nor through the interaction with the AI as the material entity.

In comparison, *strong* sociomateriality assumes that the social entity (human) and the material entity (AI) are formed only through enactment. This is in line with relational ontology, another philosophy of

sociomateriality that advocates that the social and material entities exist only in the context of that relationship, given that they are “inherently inseparable” (Cecez-Kecmanovic et al., 2014, p. 809). This view is supported by – among others – Barad (2003), Latour (1992), and Orlikowski (2007; Orlikowski and Scott, 2008), and draws on ANT, in particular, to focus on how the power of interactions depends on the association drawn by humans (Baron and Gomez, 2016; Contractor et al., 2011; Law, 1992). A *strong* form of the notion *performativity* is present when managers act, for instance, as role models, influencing not only employee ethical behavior but also their ethical ideology (*virtue ethics*) regarding interacting with AI. Further, changes in an individual’s ethical ideology can depend upon various factors, such as influences through the interaction with AI as material entity or other influences originating from the social subsystem. Hence, using the *weak* and *strong* notions of sociomateriality allows us to better explain the strength of the interdependencies between the entities (social and material entity) that are arranged in the sociotechnical system.

To sum up, by differentiating between *weak* and *strong* notions of sociomateriality, we aim to provide a holistic and systematic view of the dynamics between AI (material entity), employees (social entity), and other human and non-human objects in the *social subsystem* – all incorporating the *virtue ethics* and *duty ethics* on an individual or collective level to some extent – that are arranged in an organization as sociotechnical system. This helps us to answer our research question regarding how we can understand and further explain ethical management of human-AI interaction from both the perspective of AI interacting with humans (SQ1), and that of humans interacting with AI (SQ2), as it allows us to outline the interdependencies of *virtue ethics* and *duty ethics* in the sociotechnical system. (Appendix A provides an overview of *weak* and *strong* notions of sociomateriality based on Jones (2014) and adapted toward our context of human-AI interaction.)

3. Methodology

Following the classification of literature reviews of Paré et al. (2015), we aimed to conduct a *theoretical review* grounded in a broad literature search and qualitative content analysis of relevant articles to elaborate and expand knowledge on the ethical management of human-AI interaction through the lens described above. Our research design comprises two stages: first, a *conceptual stage* (described in section 3.1) in which we developed deductively a conceptual framework to structure our review and aid in theorizing; and second, a *review stage*

(described in section 3.2) encompassing the qualitative analysis of the literature identified based on the search protocol. Our analysis goes beyond synthesizing the literature by extending the main constructs from the conceptual stage and further examining their relationships through the notions of sociomateriality, which is the theorizing part of our review. In the following paragraphs, we elaborate on the two stages.

3.1. Conceptual Stage

The aim of this stage is to derive a conceptual framework *deductively* to guide the theorizing part of our *theoretical review* in the second stage. We establish a deductive foundation that can be extended through the theorizing part of the review, adopting an *integrating* type of conceptualizing by following the guidelines for conceptual work proposed by MacInnis (2011). Integrating involves the synthesis of main constructs – ethics, human-AI interaction, and strategic management in our work – by taking what is known so far and transforming it into a novel structure (here integrating main constructs in sociotechnical system theory as conceptualized by Leonardi (2012)) by looking at the connections between all relevant constructs (here through the lens of sociomateriality) to gain a holistic perspective (MacInnis, 2011). It is important to note that as frameworks aim “to integrate” main constructs to define boundaries (Schwarz et al., 2007, p. 44), we do not claim comprehensiveness when synthesizing existing knowledge from relevant work of reference disciplines (Burton-Jones et al., 2021) for this integration. The integrative framework in the *conceptual stage* served as an “architectural plan” (MacInnis, 2011, p. 138) to guide the *review stage*. To provide an insightful framework with desirable quality, we followed the criteria of assessment of Schwarz et al. (2007). Appendix B details how we corresponded to these criteria. Our steps to develop the framework as an *integrating type* of conceptualization are described below.

First, we used our main constructs “ethics,” “human-AI interaction,” and “strategic management” as keywords to search for relevant literature in the databases listed in Appendix C. To identify existing work, we also used synonyms and similar concepts, such as “human-AI hybrids” (e.g., Rai et al., 2019) for the main construct of “human-AI interaction” and “AI governance” (e.g., Shneiderman, 2020) for the main construct “ethics.” In our search, we used database filter-options to show the “most cited” articles (if available) and/or to sort articles with respect to their relevance. We also searched for current literature reviews summarizing the

academic discourse on our main constructs. Our aim was to identify the most relevant articles in the academic discourse.

Based on the most relevant articles that addressed our main constructs, we conducted a backward search to check the references those articles used for their theoretical foundations (Weber, 2012). In particular, we checked whether there were similar or often-used references, following a snowball sampling strategy.

All articles identified as being highly relevant to our topic under research were used to identify and synthesize subconstructs that we employed in the next step to explain our main constructs in more depth. We identified 13 articles relevant for the synthesis of the subconstructs “substitution,” “augmentation,” and “assemblage” as different types of human-AI interaction; 16 articles remained for the synthesis of subconstructs representing principles of “ethics” (in specific, 9 articles representing business ethics and 7 articles representing AI ethics, being synthesized together as categories of “human-AI ethics”); and 9 articles remained for the synthesis of the subconstructs “management levels” (n=4) and “management process” (n=5) of the main construct “strategic management.” Appendices M–O provide an overview of the identified literature we used to derive our subconstructs.

We next focused on literature that contributed directly to our research process and goals. Following Burton-Jones et al. (2021), we limited our analysis to the subconstructs that are focal to the understanding of our research subject; we focused on literature contributing to our focus on the “strategic management” of human-AI interaction; and we used the subset of articles resulting from the snowball sampling that were most relevant to our theory development objective, as it is important to “identify the set of manuscripts that contribute to a theory and distinguish them from those that cite the theory for other purposes and be able to safely ignore those” (Larsen et al., 2019, p. 892). This narrowed down the literature to those subconstructs that are most relevant to establishing an *integrating type* of conceptualization: 8 articles remained to explain the types of human-AI interaction as subconstructs of the main construct “human-AI interaction”; 9 articles remained to explain the subconstructs representing principles of human-AI ethics of the main construct “ethics”; and 8 articles remained to explain the subconstructs of management levels and processes belonging to the main construct “strategic management.”

We then integrated our main constructs and their subconstructs into Leonardi's (2012) conceptualization of *sociotechnical systems*. To avoid "poor construct conceptualization" that could lead to misspecifications during theory development (MacKenzie, 2003, p. 324), we provided explanations for how the subconstructs are interrelated and how we embedded them into Leonardi's (2012) sociotechnical system. This resulted in a conceptual framework (see section 4) that served as the foundation for our *theoretical review*, in which we outlined how the notions of sociomateriality help us systematically understand ethics in human-AI interaction.

3.2. Review Stage

We began the review stage by gaining initial insights from a preliminary literature review conducted between April and November 2021 and based on the terms from the main constructs detailed in the *conceptual stage*: types of human-AI interaction, ethics, and strategic management. The research team discussed the results and undertook conceptual mapping exercises to scope the literature in relevant fields. This process served as a stage-gate to the review process and yielded the results required for developing the search strategy. We adapted our framework based on the results of this preliminary literature review, relying on iterative rounds and leading to our research design, thus contributing to the quality of the outcome of theorizing (Rivard, 2021).

Next, we followed the steps suggested by Templier and Paré (2018) for conducting a *theoretical review*. We assessed the seven databases that index scientific publications relevant to our review to ensure broad coverage of interdisciplinary literature: *ACM Library* (ACM), *AIS Electronic Library* (AIS), *Business Source Ultimatum via EBSCOhost* (EBSCO), *IEEE Computer Society Digital Library* (IEEE), *JSTOR*, *ScienceDirect* (SciDir), and *Web of Science* (WoS). The databases cover major academic journals from the domains of Information Systems (IS), Management and Organizational Science, and ethics (see Appendix C). We manually checked leading conferences and journals that are not included in these databases (see Appendix D).

We conducted a search based on the keywords derived from the first stage (Kitchenham and Charters, 2007). The primary keywords are derived from the constructs of our conceptual framework: the types of human-AI interaction and ethics in the context of strategic management. In addition, we adopted four abstract search terms that refer to the lines of research relevant to our work: "AI," "strategy or management," and "ethics" (see Appendix E). This increased the number of potentially relevant results, allowing us to gain a broader picture of

the current literature (Boell and Cecez-Kecmanovic, 2015). We included articles irrespective of their publication date.

The search was carried out in June 2022 and updated in January 2023. It yielded 10,009 hits (9,136 hits in the main search and 873 hits in the update) across all databases without the use of database filter options; 6,904 hits (6,058 hits in the main search and 846 hits in the update) across all databases with the use of database filter options (see *format screening*); and 6,377 unique hits (5,688 hits in the main search and 689 hits in the update) after the elimination of 527 duplicates (n=370 in the main search and n=157 in the update). We added 1 hit through a manual search in the five leading conferences and journals not covered in the databases. In total, we received 6,905 overall hits (with database filter options); 6,378 unique hits remained for the subsequent manual screening processes.

In the next step, we applied exclusion and inclusion criteria through *format*, *quality*, and *relevance screenings*. In the *format screening*, we applied several exclusion criteria to remove non-scientific literature, incomplete articles, and gray literature (see Appendix F). We had already applied format criteria through database filter options whenever possible (see preceding step) and thereby removed 3,105 articles. Since this option is not available for all outlets, the format criteria also had to be checked manually (see Appendix F). In total, 2,379 articles were eliminated manually due to format criteria, and 3,999 articles remained.

We conducted a further *quality screening* based on various established rankings to restrict our review to high-quality publications that define the state of the art and to reduce the set of papers considered to a quantity appropriate for qualitative content analysis (see Appendix G). During the preliminary literature review, we identified several ethics journals that are important to the community but not present in any ranking. As an exception to the *quality screening* process, we included these journals to achieve a balance between interdisciplinary research streams and advance diversity in developing theory (Markus and Rowe, 2021). At the end of the *quality screening* 1,858 articles were removed and 2,141 remained.

Finally, we performed several rounds of *relevance screening* to increase the quality of our results. In the first round, we chose a small subsample of the articles (n=133) randomly for the parallel training purposes of the coders. The relevance criteria were revised through subsequent discussion. We used Cohen's Kappa,

following Cohen (1960), to measure interrater agreement and increase our review reliability, obtaining a score of $\alpha=0.81$, which indicates a strong agreement between coders (McHugh, 2012). Next, each of the coders coded an equal share of the remaining articles. In a third round, the articles were exchanged between the coders, thus ensuring that codes were double-checked to reach a strong level of reliability.

The *relevance screening* process was based on the full content of each paper. To be deemed relevant, each paper had to meet a set of three criteria. First, articles had to provide contributions regarding ethics in the context of human-AI interaction. Hence, for example, articles that analyze AI design principles rather than focusing on the human-AI interaction were excluded; while such articles may answer questions about how we might understand ethical AI development, they do not typically address ethical management of AI interacting with humans (see SQ1). Second, the AI-based technology mentioned in the articles had to carry out a task that typically requires human cognition, following the definition of AI based on Rai et al. (2019). Third, to be considered relevant, articles must have included ideas on the management of human-AI interaction in an organizational context (especially due to SQ2, which aims at understanding the management of humans interacting with AI). See Appendix H for a complete list of criteria. In the end, 2,121 articles were excluded, and 20 articles remained.

Since a keyword-based search may not identify all relevant articles, we also conducted a *backward and forward search*. For the *backward search*, we checked the sources cited in each of the 20 articles that remained in our final sample after the keyword search in databases and manual search (keyword sample). In addition, we conducted a *backward search* and *forward search* for the previously excluded literature reviews (review sample) and research-in-progress articles (RIP sample). For the *forward search*, we looked for articles that refer to articles of the three samples: keyword, review, and RIP. The articles identified through the *backward* and *forward searches* were further checked against the inclusion and exclusion criteria described above. The remaining articles were subjected to further rounds of *backward* and *forward searches*, a process repeated through several rounds until it no longer yielded relevant articles not already present in the dataset (see Appendix I). This added another 22 articles to our final sample.

Overall, the systematic literature search yielded a *final sample* of 42 relevant articles meeting the inclusion criteria for format, quality, and relevance across all databases and backward and forward searches (see Appendix J). Before downloading the PDFs of our final sample into the software MAXQDA for qualitative analysis, we structured the sample's information as *meta information* to provide a descriptive overview of our final sample (see Appendix K). Meta information describes the articles based on their main characteristics (e.g., research domain, method) that differentiate between different article types.

Figure 1 is an overview of the search process based on the steps suggested by Templier and Paré (2018).

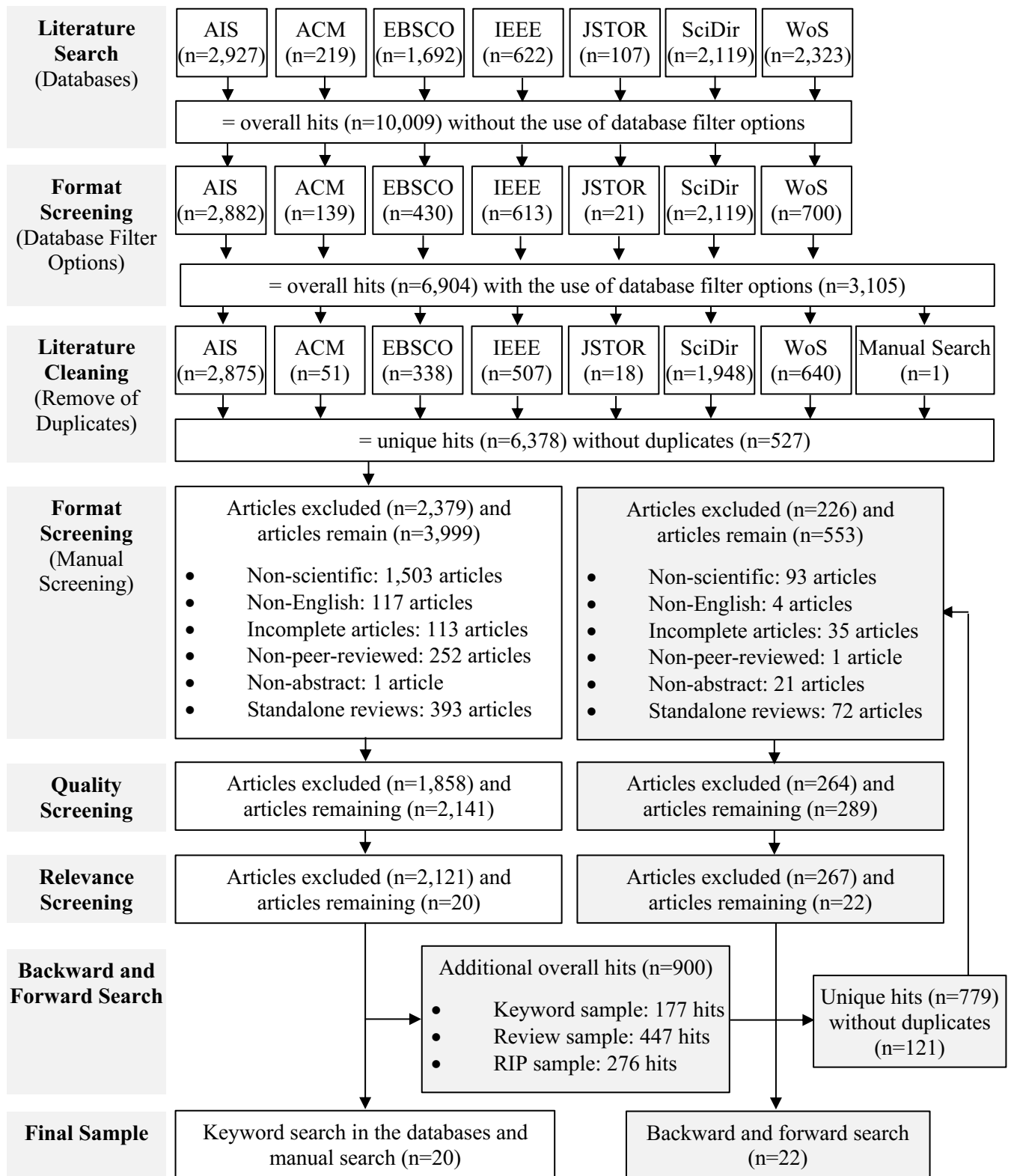


Figure 1: Search Process, based on Templier and Paré (2018)

For the *qualitative analysis* of the final sample of articles, we structured the content of relevant articles according to the constructs comprising our deductively derived framework (Mayring, 2014, pp. 95–98). To

guarantee consistency during the coding process, we used a codebook that details all deductively derived codes in terms of definitions, anchored examples, and coding rules for each category, as suggested by Mayring (2014, p. 95). The codebook was developed by all coders, and its reliability was improved through discussions among the coders and repeated coder training. Appendix L shows the coding scheme. The deductive analysis reflects the *extend strategy* of “Next-Generation Theorizing” of Burton-Jones et al. (2021) by increasing our understanding of the transforming nature of human-AI interaction as the changing phenomenon through the use of the notions of sociomateriality in the context of sociotechnical system theory (as well-established theories), and by extending those theories.

4. *Conceptual Framework*

The following pages explain the three main constructs (ethics, human-AI interaction, and strategic management) and their subconstructs that are relevant to increasing our understanding of ethical management of human-AI interaction. We further integrated them into Leonardi’s (2012) conceptualization of sociotechnical systems to demonstrate initial interdependencies. This process resulted in a deductively derived conceptual framework (see Figure 2) that serves as a foundation for the theorizing part of our *theoretical review* and helps us to better reflect upon the strategic management of human-AI interaction from both perspectives: managing AI interacting with humans (SQ1), and managing humans interacting with AI (SQ2).

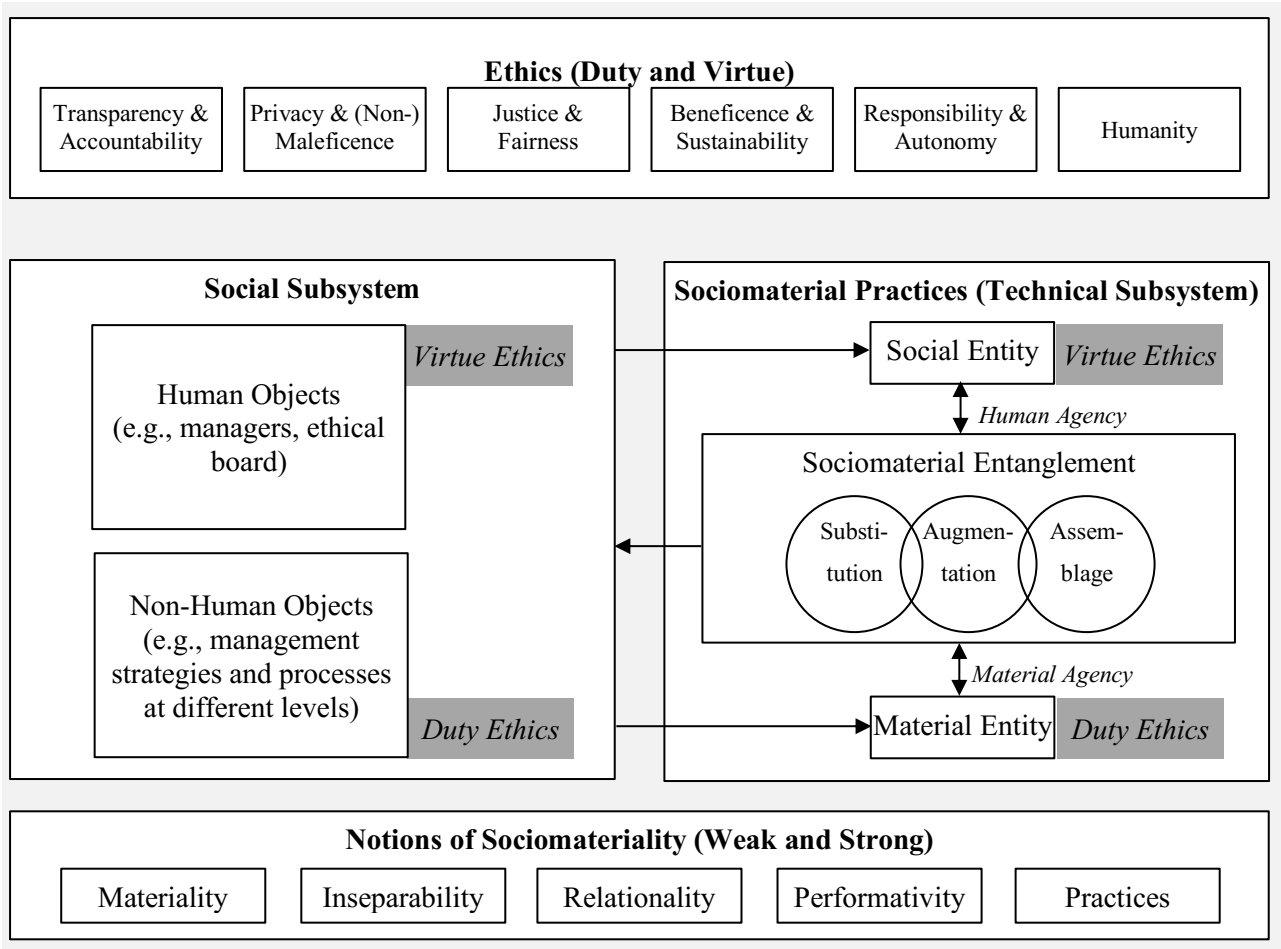


Figure 2: Conceptual Framework of Ethical Management of Human-AI Interaction

As Figure 2 illustrates, we identified three main constructs as relevant to our research context:

- i) The types of human-AI interaction (*substitution*, *augmentation*, and *assemblage*) are part of *sociomaterial entanglement* and are described further in section 4.1.
- ii) Ethics (especially *virtue ethics* and *duty ethics*) can be part of either the *social subsystem*, representing ethics on a collective level, or *sociomaterial practices*, representing ethics on an individual level. This is detailed in section 4.2.
- iii) Strategic management represents the context in which human-AI interaction can be managed and is part of the *social subsystem* (see section 4.3).

The arrows in Figure 2 between the *social subsystem* and *sociomaterial practices* are based on Leonardi (2012), who promotes the substantialist ontology (*weak sociomateriality*) and suggests that the *social subsystem* and *sociomaterial practices* are mutually dependent. *Sociomaterial entanglement* may affect the *social*

subsystem, while processes within the *social subsystem* “can shape future patterns of imbrication, [that] can in turn bring changes to an artifact’s materiality or a person’s intentionality” (Leonardi, 2012, p. 43). Even though we adopt this ontological view in the conceptual framework of Figure 2, we follow the recommendation of Jones (2014) in our second stage, when theorizing based on the literature review, to achieve a more differentiated analysis of *weak* and *strong sociomateriality* through the notions of sociomateriality.

4.1. How the Types of Human-AI Interaction Refer to Sociomaterial Entanglement

As Figure 2 outlines, human-AI interaction resembles *sociomaterial entanglement* embedded in *sociomaterial practices*. Since increased capabilities of AI lead to changing natures of sociomaterial entanglement, we need to differentiate between different types of human-AI interaction. It can be classified into three archetypes (see Appendix M): *substitution*, *augmentation*, and *assemblage*.

Substitution refers to AI that replaces humans in performing a task (Rai et al., 2019). Strictly speaking, no specific human-AI interaction concerns the specific task the AI conducts. Hence, ethical management must focus on the conditions surrounding these tasks, as well as upstream and downstream processes.

Augmentation focuses on improving each other’s abilities (Rai et al., 2019) and balancing each other’s weaknesses (Daily et al., 2017; Raisamo et al., 2019). Either the AI can improve the human’s performance (*human augmentation*; Wilson and Daugherty, 2018, p. 214), or the human can improve the AI’s performance (*machine augmentation*; Mnih et al., 2015). We feel, though, that this definition does not adequately reflect different cases of augmentation, as there may be different outcomes (positive or negative). In some cases, humans as the social entity may try to exploit the algorithm as the material entity to increase their personal benefits (Lee, 2018; Lee et al., 2015; Zhang et al., 2020), as awareness of the weaknesses of the systems surrounding the AI makes humans able to behave opportunistically (representing a rather egoistic form of opportunistic behavior). In other cases, humans may “outsmart” the AI, not for their own benefit, but rather for an overarching goal from which others benefit (representing rather an altruistic form of opportunistic behavior). In this way, one entity may try to balance the insufficient moral behavior of another in accordance with ethical norms. The ideology of the social entity does not subordinate morally correct behavior to one’s own advantage; instead, it is about weighing different moral maxims against each other (e.g., helping others vs. behaving

according to transparency standards). Thus, when considering managing *augmentation*, we need to account not only for the behavior but also the ethical intention of humans (that may be positive or negative) that guide the behavior in human-AI interaction.

Assemblage (often referred to as hybrid intelligence) suggests that combining humans and AI agents in an integrated unit to perform complex tasks leads to better outcomes through collective intelligence (Rai et al., 2019). We also suggest patterns of interpreting *assemblage* as a process of several augmentative or substitutive tasks being executed sequentially or in parallel, leading to “superior results” (Dellermann et al., 2019a, p. 19). This process view demonstrates that there can be tasks during human-AI interaction in which the AI replaces the human and other tasks in which the human or the AI is augmented by the other entity.

Fuegener et al. (2021, p. 32) state that “metaknowledge” is required to achieve “superior results” with AI by deciding which tasks should be delegated to the AI (*substitution*) and which should remain with humans (*augmentation*), given that “poor delegation decisions” can lead to “unconscious traits” and a low level of performance. Bansal et al. (2021, p. 13) state that “in an ideal team, the human and AI would have minimally overlapping mistakes so that there is a greater opportunity to correct each other’s mistakes.” This demonstrates that *assemblage* is built upon *substitution* and *augmentation*. Therefore, ethical management of *assemblage* needs to consider both upstream and downstream processes (resulting from the insights we derived for substitution) as well as ethical behavior and the intention of the social entity (which may be positive or negative), which we learned from the insights we derived for augmentation.

While we adopted different types of human-AI interaction as defined by Rai et al. (2019), we aim to create awareness that ethical management of human-AI interaction cannot be restricted to a specific task and, thus, to a specific type of interaction. Consequently, the process of entanglement in the context of *sociomaterial practices* cannot be strictly classified according to these types of human-AI interaction. Instead, these archetypes should be understood as tools that help us make the process view and its effects on ethical management more tangible. We refer to this in the second stage when presenting and discussing the review results.

4.2. How Ethics is Considered as Part of Both the Social Subsystem and Sociomaterial Practices

As Figure 2 illustrates, ethics on an individual level can be subordinated to the social entity and the material entity being part of *sociomaterial practices*, but also on a collective level to human objects and non-human objects being part of the *social subsystem*. Below, we deepen our understanding of ethics and its subconstructs to understand the differences in ethical considerations that can be drawn depending on where in sociotechnical systems we are talking about ethics.

In general, ethics can be seen as “the capacity to think critically about moral values and direct our actions in terms of such values.” (Churchill, 1999, p. 259). Hence, ethics is thinking not only about moral values as “the intention of the person making the decision or performing the act,” but also about “the consequences of ... an action or decision” (Hosmer, 1987, pp. 321–322). In this work, we refer to *virtue ethics* when referring to a person’s ethical intention, which is guided by a person’s ethical ideology and characterized by the person’s moral values and assumptions determining the intrinsic motivation to behave ethically or not (Eitel-Porter, 2021; Flathmann et al., 2021; Gandz and Hayes, 1988). Beyond *virtue ethics* that guide ethical behavior, there are also principles, rules, norms, and standards that guide ethical behavior by obligation, which we refer to as *duty ethics* – also known as deontology (Bowie, 1998; Flathmann et al., 2021; Jones et al., 2005; Kant, 1991, pp. 66–70).

As outlined in the Introduction, ethical management of human-AI interaction on a strategic level requires consideration of both ethical perspectives, *duty ethics* and *virtue ethics*, both of which are represented in the *social subsystem* and *sociomaterial practices*. *Duty ethics*, inter alia, is part of the social subsystem, as it reflects guidelines and norms in an organization (Leonardi, 2012, p. 43). The *social subsystem* also includes the *virtue ethics* perspective since the workforce has its own ethical understanding based on collective moral values that drive ethical behavior in the organization regardless of *duty ethics*, and that might differ from the ethical ideology of any individual worker (Coldwell et al., 2008; Stevens, 1999). Both ethical perspectives are also part of *sociomaterial practices*, given that within their enactment, the AI as the material entity is designed on ethical principles guiding the material agency (*duty ethics*), and the employee as the social entity has its own inherent ethical ideology guiding human agency (*virtue ethics*) (Flathmann et al., 2021; Siau and Wang, 2020). While the *social subsystem* comprises ethical principles and guidelines on a collective level (*duty ethics*) and corporate values based on the ethical understanding of the workforce (*virtue ethics*), *sociomaterial practices* comprise both ethical perspectives on an individual level concerning a specific task.

When applying both ethical perspectives to our underlying research subject, we also have to consider established research in related domains, specifically AI ethics and business ethics. AI ethics corresponds to SQ1 (AI interacting with humans), whereas business ethics is linked to SQ2 (humans interacting with AI). Both AI ethics and business ethics are based on principles that guide ethical behavior (Floridi et al., 2018; Rezaee, 2008, p. 64) (see Appendix N). AI ethics comprises “a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies” (Leslie, 2019, p. 3). Business ethics comprise organizational norms and guidelines that are required to guide ethical behavior among organizational stakeholders, as some people in organizations may act in what could be characterized as a morally insufficient manner (Jones et al., 2005; Parker, 2002). AI ethics and business ethics correspond primarily to the *duty ethics* perspective that aims to foster ethical behavior in human-AI interaction and, to some extent, aims to strengthen an individual’s ethical understanding (what we refer to as *virtue ethics*).

To make ethical considerations in human-AI interaction more tangible, we draw upon the principles discussed in the literature on AI ethics and business ethics and derived six categories of “human-AI ethics” (for details, see Appendix P):

- i) *transparency & accountability* aim to make human-AI interaction processes easy to understand and decisions reviewable, and to ensure that processes of human-AI interaction can be investigated depending on the interests of both entities (social and material entity);
- ii) *privacy & (non-)maleficence* aim to protect personal privacy and security in human-AI interaction, avoiding misuse of AI through humans and vice versa;
- iii) *justice & fairness* aim at respecting all the interests of both entities (social and material entity) to maximize inclusion and fair treatment in all types of human-AI interaction;
- iv) *beneficence & sustainability* aim at beneficial human-AI interactions with a long-term positive impact;
- v) *responsibility & autonomy* aim to balance the power of AI and humans in their interactions and manage human-AI interaction based on the agreement among all entities (social and material entity) involved or affected by its consequences; and

- vi) *humanity* aims to maintain social values and foster social togetherness in the context of human-AI interaction

We refer to these categories of human-AI ethics to better explain the interdependencies between the *virtue* and *duty* ethics perspectives in managing human-AI interaction.

4.3. How Strategic Management Represents the Context of the Social Subsystem

In the context of information technologies (IT), a strategy concerns “the ability to identify and evaluate the implications of IT-based opportunities as an integral part of business strategy formulation” (Peppard and Ward, 2004, p. 176). Strategic management includes strategies at three levels. At the *corporate level*, strategy (corporate or global strategy) concerns the strategic planning and coordination of several substrategies (business strategies). At the *business level*, a strategy aims to gain a competitive advantage and achieve the organization’s overarching goals and vision (Karimi and Konsynski, 2002, p. 92; Ritson, 2008, p. 13; Somogyi and Galliers, 2009, pp. 11–12). The *functional level* has strategies based on different departments, such as accounting and HR (Karimi and Konsynski, 2002, p. 93; Somogyi and Galliers, 2009, p. 15). To achieve a “strategic fit” (Henderson and Venkatraman, 1999, p. 476), it is important to align these strategies with the organization’s processes.

As Figure 2 illustrates, strategic management can be interpreted as a context being part of the *social subsystem* that guides ethical behavior through strategies and processes before (preventive), during, and after (reactive) the human-AI interaction in an organization. Strategic management establishes ethical principles that provide guidance and standards for how human-AI interaction can be accompanied by efforts to manage it ethically. When managed well, AI can increase the quality of work, but poorly managed AI can increase opacity and inequality in work arrangements (Faraj et al., 2018; Jobin et al., 2019). The strategic management of human-AI interaction in an organization can be split into four processes: *planning*, *strategic*, *tactical*, and *operational processes* (see Appendix O for an overview). Given the relationship between strategic management and organizational goals (Henderson and Venkatraman, 1999; Peppard and Ward, 2004), organizations must involve their ethical understanding when formulating strategies and aligning their processes to achieve ethical management of human-AI interaction. This means that an organization’s ethical principles are represented in

certain ways through its strategies and processes – what can be referred to as AI governance (Gasser and Almeida, 2017) and corporate governance (Bonn and Fisher, 2005; Rezaee, 2008). For example, to enforce principle-based ethics, organizations may follow external guidance (Rezaee, 2008; Shneiderman, 2020).

In the analysis that follows – the second stage of our theory-developing review – we aim to provide deeper insights into the ethical management of human-AI interaction by explaining the relationships and interdependencies between all constructs and their subconstructs we introduced in this stage through the five notions of sociomateriality.

5. *Analysis and Results*

In this section, we describe and explain the main findings of our review along the five notions of sociomateriality (*inseparability, relationality, performativity, materiality, and practices*). The notions enable us to better reflect the complexity of human-AI interaction (*sociomaterial entanglement*) and explain the interrelationships with the corresponding management practices and processes (*social subsystem*) occurring in organizations as sociotechnical systems. Given that our underlying aim is to increase understanding of the ethical management of human-AI interaction through the lens of sociomateriality, we explain in this section how *virtue ethics* and *duty ethics* can constitute the ethical management of human-AI interaction corresponding to our two subquestions concerning AI interacting with humans (SQ1) and humans interacting with AI (SQ2). Hence, we concur with Jones' (2014) arguments regarding the relevance of using the potential of the notions of sociomateriality to explore sociotechnical interactions – especially when it comes to ethical considerations.

Figure 3 is an overview of how the five *weak* and *strong* notions of sociomateriality can be integrated into the conceptual framework we derived in the conceptual stage (see section 4) based on the main findings of our literature review.

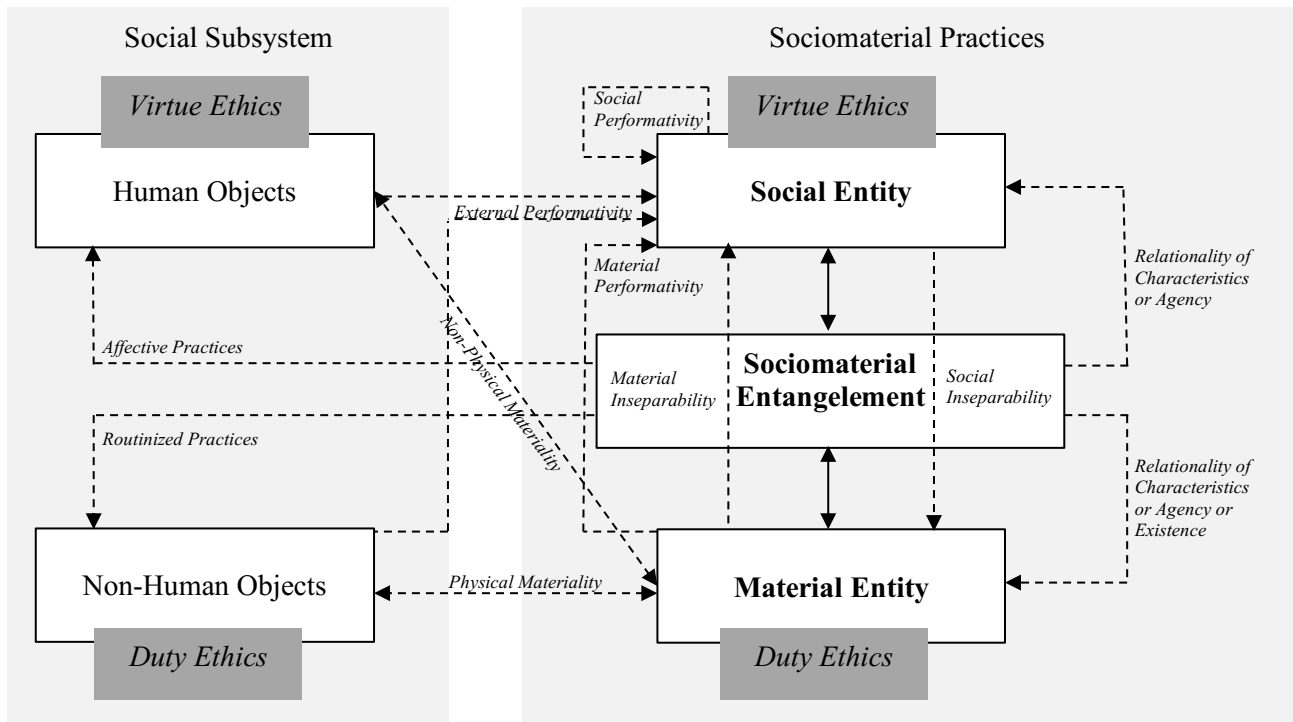


Figure 3: Notions of Sociomateriality Determining Ethical Management in Sociotechnical Systems

Although Figure 3 represents the two ethical perspectives (*duty ethics* and *virtue ethics*) assigned to different entities (social or material entities) or objects in the sociotechnical system, there are no clear dividing lines between those two perspectives. For instance, Coombs et al. (2021, p. 4) provide an example that demonstrates these mutual interdependencies of (*virtue* and *duty*) ethics, given that “[f]or machines to be valuable overall, they must support higher-level, more socially beneficial thinking and discussion by humans” – and hence it is important to integrate *virtue ethics* in the AI’s logic. Thus, the machine’s behavior is based on “the lens through which [AI designers] look at and reason about the world is shaped by [their] conscious and, more importantly, unconscious values and assumptions” (Ehsan and Riedl, 2020, p. 12). Following this argument, the AI also inherits the developer’s *virtue ethics* to some extent, resulting in programmed rules (which we refer to as the AI’s *duty ethics*).

In the following, we detail the interweaving of all constructs constituting the ethical management of human-AI interaction and their interdependencies by investigating each notion of sociomateriality and introducing the subforms outlined in Figure 3 based on the results of our literature review.

Inseparability: How interdependencies between the social entity and the material entity affect ethical management of human-AI interaction

The notion of *inseparability* focuses on the interaction (moment of “imbrication”) of the social and material entity (*sociomaterial entanglement*). We suggest distinguishing between two subforms of *inseparability*:

- i) *Social inseparability*: the material entity depends on the social entity within the sociomaterial entanglement (e.g., decisions of humans build the material entity and how it operates). In a *strong* form, this can be aligned with *machine augmentation*, whereas in a *weak* form, it can be linked to *substitution*.
- ii) *Material inseparability*: the social entity depends on the material entity within the sociomaterial entanglement (e.g., the material entity determines working practices). In a *strong* form, this can be aligned with *human augmentation*, whereas in a *weak* form, it can be linked to *substitution*.

The following two sections provide examples from our literature review to bolster our suggestion to differentiate between the two subforms and demonstrate how they shed light on the interdependencies between the social and material entity when directly entangled (wherefore we also relate the subforms to the types of human-AI interaction).

Weak social inseparability (the material entity operates without or with less human intervention) is linked to **strong material inseparability** (the work practices of the social entity are highly dependent on the material entity): In the context of recruitment, the AI (material entity) that can, for instance, reject job applicants without human intervention (Acikgoz et al., 2020) indicates *weak social inseparability*, as the AI decides independently of the recruiter (social entity); this indicates *substitution*. However, in the “second stage of staffing,” the recruiter (social entity) depends on the AI’s decision regarding whom to invite for a job interview, demonstrating that certain tasks cannot be completed without the material entity (*human augmentation*) and, thus, indicating *strong material inseparability*.

Strong social inseparability (the material entity cannot interact or operate without the social entity) is linked to **weak material inseparability** (the material entity falls out of the awareness of the social entity in work practices): In the context of loan decisions, the loan consultant – as the social entity having expertise on loan

decisions – might not depend as much, compared to a non-experienced employee, on the AI-based systems (material entity) providing decision criteria for loan acceptance or rejections. Strich et al. (2021, p. 316) offer an example from an interview with a former loan consultant: “I know how to deal with the customer if the system is offline. It may sound mean, but the new colleagues would have a big problem, because they are only used to working according to the predefined path [given by the AI-based system]. In addition, if something beyond that happens, they would be helpless.” In the example, the former loan consultant (social entity) does not depend on the AI (material entity), which indicates *weak material inseparability*. At the same time, the AI-based system depends on previous inputs of the social entity: the “AI system uses past data from similar customers profiles to predict future customer needs” (Strich et al., 2021, p. 315). That engenders *strong social inseparability* and refers to *machine augmentation*.

However, the level of interdependence can vary over time given that – following the aforementioned example of integrating an AI system for loan decisions – new colleagues are not experienced loan consultants (social entity). They may not be able to work without the AI (material entity) due to their lack of knowledge (Strich et al., 2021), which demonstrates a *strong material inseparability* and *human augmentation*. These interdependencies over time confirm the relevance of a process view when it comes to different types of human-AI interaction (see section 4), and so it is important to differentiate between several forms of *inseparability*. In the following, we elaborate on these four different forms of the notion *inseparability* and highlight how they affect the two different ethical perspectives (*virtue* and *duty*) in the management of human-AI interaction.

Beginning with *weak social inseparability*, we recognize that this form is characterized by *transparency & accountability* issues (referring to the *duty ethics* perspective), given that “employees [note] a lack of control over the interaction and outcome” (Rix and Hess, 2022, p. 9) due to the complexity of the algorithms (material entity). This leads to “highly unpredictable” results (Mayer et al., 2020, p. 248) and “heightens [individual] perceptions related to the creepy nature of the [AI-based] technology” (Rajaobelina et al., 2021, p. 2351). Hence, a “clear capability to explain intent and process” is required (Someh et al., 2022, p. 151) to enhance the decision confidence of humans having to work with the AI’s outcome (Liao et al., 2020). Hence, *duty ethics* is especially relevant when the AI (material entity) operates without human interventions (thus, no ability to involve someone’s *virtue ethics*).

By contrast, ***weak material inseparability*** is linked to the *virtue ethics* perspective of humans since “usually any final decision ... is still taken by humans – and in many cases, AI suggestions ... are not implemented” (Klumpp and Zijm, 2019, p. 269). For instance, in the context of using an AI agent to predict a project’s success or failure, “humans are using the agent as a filter [for] a first check” (Basu et al., 2021, p. 8). Not following the AI suggestions can be a “sign of resistance” or result from the human assuming the possession of “better knowledge” compared to the AI (Klumpp and Zijm, 2019, p. 269). Yet, the interaction can also promote ethical behavior when the material entity and the social entity operate as separate units (*weak material inseparability in substitution*). This is especially the case when we look at *beneficence & sustainability* from a *virtue ethics* perspective, given that AI can help create favorable work environments and increase work-life quality (Fuegener et al., 2021; Rix and Hess, 2022). For instance, the AI (material entity) can take over “disliked and routine tasks,” leaving employees (social entity) to “focus on more value-creating and experimental tasks that correspond with their personal interests” (Rix and Hess, 2022, p. 7) and that are more related to “creative thinking” (Malik et al., 2021, p. 344). This represents a potential to enforce beneficence between humans and AI, interacting mainly independently of each other. However, as the interaction between humans as social entities is not mediated by the material entity, there may be other mechanisms required to ensure correct ethical behavior of both entities.

Concerning ***strong social inseparability***, we recognized the increasing importance of the *virtue ethics* perspective, as the material entity cannot operate adequately without the social entity. This can be linked mainly to the issue of *humanity*. Therefore, it is essential to align human values (*virtue ethics* of the social entity) with algorithmic values (*duty ethics* of the material entity) (Mayer et al., 2021). In this way, ethical guidelines (*duty ethics* of the *social subsystem*) can help create a “collective understanding of AI ethics in the corporation” (*virtue ethics* of the *social subsystem*) (Mayer et al., 2021, p. 6). The ethical guidelines of the *social subsystem* can moderate the alignment between humans’ *virtue ethics* and the AI’s *duty ethics*.

It is important, however, to include human judgment and the ability to create situational awareness in human-AI interaction (e.g., Malik et al., 2021, Park et al., 2021, Someh et al., 2022). For instance, one interviewee of Someh et al. (2022, p. 153) argues that “[i]f something is going against a person, we wanted humans to intervene and make final judgment as opposed to a machine making the judgment.” Lee et al. (2015,

p. 1610) demonstrated that there can arise “many complications ... when one relies too heavily on quantified metrics without deeper considerations of their meaning and nuances.” Hence, to foster humanity and reduce the emotional burden during human-AI interaction, it is important to align AI’s *duty ethics* with human *virtue ethics*.

In comparison, *strong material inseparability* is often linked to ethical issues from the *virtue ethics* perspective regarding *responsibility & autonomy*. For instance, the “loss of critical thinking and knowledge” can become an issue for organizations in the long term, given that “employee’s awareness is severely impaired or even nonexistent” (Mayer et al., 2020, p. 249). Thus, they simply transfer their responsibility to the AI, which leads to employees not justifying decisions (Strich et al., 2021). The loss of autonomy of the social entity further spurs *strong material inseparability* and can even increase when the AI restricts humans’ critical thinking about decisions through a programmed carrots-and-sticks mechanism (*duty ethics*). For instance, the platform Uber, which automates “the exercise of managerial control, uses “material benefits and privileges through Uber Pro, as well as symbolic badges” following strict instructions of the algorithm without questioning it (Pregenzer et al., 2021, p. 9). It may, as a consequence, might punish people who criticize the decisions, which may lead to an increased loss of autonomy, leading to unintended negative consequences. Therefore, it is “essential that an individual who may be subject to a purely automated decision has the right and power to challenge the decision” (Polyviou and Zamani, 2022, p. 9). This is why it is important to establish not only adequate principles guiding human-AI interaction but also strengthen ethical understanding and behavior by enabling individuals to criticize AI’s behavior.

To sum up, the direction of interdependencies (whether we have *weak* or *strong material inseparability* or *social inseparability*) may also influence ethical potentials or ethical issues that arise in human-AI interaction.

Relationality: How sociomaterial entanglement affects the existence of the social entity and the material entity or their characteristics, and thus, the ethical management of human-AI interaction.

Similar to the notion *inseparability*, *relationality* is also related to sociomaterial entanglement, but it focuses on the outcome of the human-AI interaction process as it describes how social and material entities, their existence, and their agency change through the entanglement. Since *relationality* focuses on the outcomes of human-AI interaction, it can be observed in all types of human-AI interaction. Yet, the same type of human-AI

interaction can lead to different outcomes regarding the characteristics of the social or material entity, its agency, and the existence of other objects. We suggest distinguishing between three subforms of *relationality*:

- i) *Relationality of characteristics*: certain characteristics of the entities (social or material entity) arise through sociomaterial entanglement (e.g., capabilities, attributes, forms).
- ii) *Relationality of agency*: sociomaterial entanglement changes the agency of the entities (material or social; agency is the capacity to carry out an action following a cognitive process linked to the entity's intentions).
- iii) *Relationality of existence*: the existence/presence of an entity is the product of sociomaterial entanglement.

The following two sections provide examples to bolster our suggestion to differentiate between the different subforms. After we introduce the difference between relationality of the social and the material entity, we explain the three subtypes by outlining how the outcome of *sociomaterial entanglement* can affect social or material entities.

Beginning with *relationality of the social entity*, the implementation of an AI-based application system may remove “consultants from decision making” and thus affects “their professional image and external perceptions of their role” (Mayer et al., 2020, p. 248), which can be linked to *strong relationality of agency* (meaning that the agency of the social entity changes through the interaction). In the long term, the consultant’s decrease in decision-making power can further lead to a change or decrease in expertise (Mayer et al., 2020), which can be linked to *strong relationality of characteristics* (the characteristics of the social entity change through the interaction). With respect to *relationality of existence*, new roles for the social entity can arise, given that “[h]uman-AI partnering will also require workers to manage shifts between micro-identities (i.e., new roles that are enacted in relation to the needs and requests of [AI-based digital assistants], and traditional human-facing roles)” (Cranefield and Doyle, 2022, p. 13). This is supported by Grønsund and Aanestad (2020, p. 1), who stated that “[i]n contrast to replacing human work, the emergent configurations required new roles and redistribution of extant expertise to augment and improve the accuracy of the algorithm.” Due to these interdependencies, relationality of existence manifests primarily in *strong* relationality.

When it comes to *relationality of the material entity*, we can see that through “human-machine collaboration, ... machines become increasingly intelligent” (Jia et al., 2022, p. 388), as for example, feedback loops during the interaction “improve AI performance” (Liao et al., 2020, p. 6). This increased intelligence can be associated with a change in the *characteristics* of the material entity. The relationality can be either weak or strong, with the characteristics changing either through the human-AI interaction (indicating *strong relationality*) or existing independent of the human-AI interaction (indicating *weak relationality*). Concerning *relationality of existence*, Coombs et al. (2021, p. 5) state that the human-AI interaction may also result in “[n]ew roles for machines through AI,” which indicates that the new roles’ *existence* depends on the interaction representing a *strong* form of *relationality of existence*. Another example illustrates *relationality of agency*. When failure is an outcome of human-AI interaction, individuals sometimes “attribute [their] agency to the technology when they lost the game, compared to the success scenario” (Jia et al., 2022, p. 393). In doing so, the social entity transfers the agency of being responsible for the outcome to the AI in either a *strong* or a *weak* form.

These examples demonstrate that there are different outcomes of human-AI interaction that can be represented through the three forms of *relationality*. In the following, we elaborate on the three different forms of the notion *relationality* and highlight how they affect the two different ethical perspectives (*virtue* and *duty*) in the management of human-AI interaction.

With respect to *relationality of characteristics*, we can observe changes of the capabilities of humans (social entity). For instance, if AI is responsible for the decisions being made in a process, it can lead to a perceived loss of competencies (Mirbabaie et al., 2022, p. 86) and the perceived loss of reputation of the social entity (Mayer et al., 2020). If this perceived loss of reputation by the social entity arises through the human-AI interaction, it is linked to *strong relationality of characteristics*. Suppose the AI gains a higher reputation through the shift of responsibility from humans to AI (representing changes in the material entities’ characteristics). In that case, it is linked to *strong relationality of characteristics*.

In contrast to a reputation loss of humans, implementing AI-based systems in an organization can lead individuals to have to “upgrade their understanding of the moral and ethical implications of these new systems”

(Watson et al., 2021, p. 469). This leads to increased capabilities of the social entity and thus demonstrates a *strong* form of *relationality of characteristics*. In this case, the individual's ethical behavior (*virtue ethics*) is also influenced by the principle-based behavior of the AI (*duty ethics*). This influence can be either positive or negative, given that the human-AI interaction can either increase or decrease human capabilities depending on how humans "counter the basic need for autonomy" (Pregenzer et al., 2021, p. 12). Ethical considerations regarding the effects of transferring capabilities from humans (social entity) to AI (material entity) arise, such as "What is the trade-off for people? What do they gain from adapting their skills, and what do they lose?"—and thus affects ethical relations of employees' *virtue ethics* (Coombs et al., 2021, p. 5). In comparison, *weak relationality of characteristics* implies that the social entities continue to do "what they are best at while AI shows objective results." (Park et al., 2021, p. 9). Hence, in the *weak* form, there are no effects between the ethical perspectives of social and material entities.

When it comes to *relationality of agency*, we can observe in the study of Cranefield and Doyle (2022) how its *strong* form can be related to ethics. The authors discovered that teachers interacting with an intelligent writing system while writing reviews of their students' work changed their agency when the system monitored and evaluated their writing (Cranefield and Doyle, 2022). The teachers shifted their focus to improving their own writing quality rather than evaluating their students' writing. Thus, the way individuals change their agency also affects their behavior in interacting with others (Mayer et al., 2020). This shows how the material entity's *duty ethics* curtails the human agency, which leads to changes in the *virtue ethics* of the social entity.

With respect to *relationality of existence*, the increased autonomy and "fast evolving capabilities" of technology lead to new roles of the AI, such as "new forms of taxing robots [that act] in a way that society once would tax humans" (Coombs et al., 2021, p. 5), representing *strong relationality of existence* of the material entity. However, human-AI interaction may also change the roles of individuals as social entity (also representing a *strong* form of *relationality of existence*). For instance, in the context of data analysis, the human-AI interaction requires new roles for data scientists as tasks switch from manual data classification toward training the AI and "providing auditing process with a ground truth against which the algorithmic outcome [is] compared." (Grønsund and Aanestad, 2020, p. 14). In doing so, the individual as the social entity must evaluate the correctness of the AI (material entity) based on its inherent understanding of correct behavior (*virtue ethics*).

To sum up, (positive or negative) outcomes of human-AI interaction result in different changes for the AI (material entity) and the employee (social entity) – especially in the *strong* subforms of the notion *relationality*. Managed well, these changes in an entity’s characteristics, existence, or agency can compensate for the ethical concerns that have arisen through the entanglement or strengthen ethical potentials. Managed poorly, they can enhance ethical issues.

Performativity: How boundaries, limitations, and perceptions of the social subsystem, the social entity, and the material entity affect the social agency and, thus, the ethical management of human-AI interaction

Performativity is related to boundaries, limitations, and perceptions of the social subsystem, the social entity, and the material entity affecting the agency of the social entity. Depending on which entity (material entity or social entity) or objects (human and non-human objects) of the *social subsystem* create the perceptions about AI, we differentiate between three different forms *performativity*:

- i) *External performativity*: the objects from the social subsystem determine the reality from outside (e.g., control mechanisms, guidelines, and norms influence the social entity’s perceptions).
- ii) *Social performativity*: boundaries of the social entity create perceptions of an own reality (e.g., humans’ mental models influence their perceptions).
- iii) *Material performativity*: the underlying models of the material entity create perceptions of humans’ reality (e.g., data and algorithmic limitations influence the social entity’s perceptions)

The following section provides examples based on the literature review to bolster our suggestion to differentiate between the three subforms of *performativity* and explain how the notions shed light on how humans perceive AI based on the boundaries, limitations, and perceptions of their own entity, the material entity, and the objects from the *social subsystem*.

In the case of *external performativity*, public perceptions of AI may be influenced by external factors, such as media that often induce “dystopia surrounding AI” (Rix and Hess, 2022, p. 8). Employees may perceive working with AI as competing against it rather than cooperating with it (Rix and Hess, 2022, p. 8). How the social entity trusts its organization and perceives the social influence of its colleagues (*social subsystem*) can

also influence humans' perceptions regarding AI (Figueroa-Armijos et al., 2022). Depending on how much the individual is influenced by media (*social subsystem*), the *external performativity* can be either *weak* or *strong* – meaning that either there are limiting mechanisms of the social subsystem that affect human perceptions of AI (*strong*) or not (*weak*).

Concerning *social performativity*, we recognized a *strong* form when, for instance, individuals perceive AI not as a system but as a “machine teammate,” as it made individuals “more tolerant of mistakes, to a level similar as to human team members” (Rix and Hess, 2022, p. 7). Not considering AI as a “machine teammate” (*weak social performativity*), by contrast, may tempt humans to “hide behind the output of the machine” and “blame the system, even though ... it is not responsible” (Rix and Hess, 2022, p. 9). This is also demonstrated by Jia et al. (2022), who investigated individuals playing games with an app that gives recommendations for the “fastest route to complete the task” (p. 391). They showed that individuals “who self-reported being less sociable tended to attribute less intention to the app when they had lost than when they had won.” Negative cognitions of humans, in particular, can lead to perceiving AI as an “inappropriate decision maker” (Bankins et al., 2022, p. 37). Hence, how humans build perceptions of AI can represent either *weak* (the boundaries of humans do not affect its perceptions) or *strong* (the boundaries of humans affect its perceptions) *social performativity*.

With respect to *material performativity*, the limitations and boundaries of AI as material entity can affect individuals as the social entity working with AI. For instance, Lee et al. (2015, p. 1608) studied Uber and Lyft drivers managed through algorithms and showed that a supply-demand control algorithm's inability to “account for feelings of unfairness” can influence drivers' behavior. Shin and Park (2019, p. 282) identified that algorithms create reality constructions of individuals by trying to “match with users' previous experience,” whereby “the more users use algorithms, the narrower their perspectives become” given that “people perceive what they want to in algorithmically recommended contents while try to ignore other viewpoints.” Hence, the multiple ways in which the social entity creates its reality based on the AI's functionalities, limitations, and boundaries can result in *weak* (limitations of the AI do not affect humans' perceptions) or *strong* (limitations of the AI affect humans' perceptions) *material performativity*.

These examples demonstrate that there are different ways in which human perceptions of AI can be influenced, which can be represented through the three forms of *performativity*. In the following, we elaborate on the three *strong* forms of the notion *performativity* (*strong material*, *social*, and *external performativity*) and highlight how they affect the two different ethical perspectives (*virtue* and *duty*) in the management of human-AI interaction. *Weak* forms of the notion *performativity* indicate that the social entity's perception of AI is not affected much by others (*social subsystem* or external factors); here, though, we focus on the *strong* form, which actually causes changes in individuals' perceptions regarding AI and, thus, may affect their behavior.

With respect to *strong external performativity*, we observe that obligatory guidelines (*duty ethics*) that “serve as decision-making support for employees in critical cases” can lead to employees who “do not have to be concerned about whether decisions are ethically justifiable from their point of view, but can rely on the guidelines in making and justifying their decisions” (Mayer et al., 2021, pp. 7–8). Thereby, human's *virtue ethics* are influenced by those guidelines from the *social subsystem*. An individual's *virtue ethics* may also be driven voluntarily by having leaders who “upgrade their understanding of the moral and ethical implications [to] successfully drive organizations forward with innovation in AI” (Watson et al., 2021, p. 469). In this way, leadership from the *social subsystem* influences the ethical ideology of employees (*virtue ethics*) and their perceptions of AI. Hence, the “social influence” of the social subsystem can affect “what individuals deem to be ethical” (Figueroa-Armijos et al., 2022, p. 11). This may also cause individuals to have ambivalent perceptions and expectations regarding AI, depending on whether they interact with an AI agent in their private lives in addition to a professional context (Hinsen et al., 2022). In their private lives, individuals may “do everything that is practical and fun” with AI agents because they possess “full acceptance,” whereas AI use in a professional context may be accompanied by concerns about being replaced (Hinsen et al., 2022, pp. 200–201). Hence, *strong external performativity* can have significant effects on an individual's *virtue ethics*, and the context of external influence (e.g., at home or the workplace) is particularly important when considering how to manage human-AI interaction ethically.

When it comes to *strong social performativity*, we recognize ethical issues regarding *humanity*. For instance, when employees that work with AI claim that “[AI] is not only the system, it is much more,” (Rix and Hess, 2022, p. 7) or when “[o]perators who worked directly with the robot regularly characterized their

relationship with the robot in collegial or personal terms, referring to the robot as their ‘work partner’ ... or ‘friend’” (Sauppé and Mutlu, 2015, p. 3617), we have a *strong* form of *social performativity* that can affect individual *virtue ethics* in a strong way. For instance, overtrust can occur when employees tend to trust the system more than they trust colleagues by ascribing greater capabilities to AI than to their colleagues (*virtue ethics*). “Individuals who are inherently more social tend to view their technological teammates in more social ways as well” and are thus more likely to attribute agency to the AI (Jia et al., 2022, p. 393). To mitigate the negative effects of overtrust, “explanations [about the AI] should be informative, instead of just convincing” (Bansal et al., 2021, p. 13).

With respect to *strong material performativity*, *responsibility & autonomy* considerations play a crucial role. For example, individuals aware of the limitations of AI may introduce their own control mechanisms to overcome those limitations as a way to preserve their autonomy (*virtue ethics*). Pregoner et al. (2021), for instance, illustrate that algorithmic-managed Uber drivers use dashcams for defense in case passengers make an accusation against them that would result in the (temporary) deactivation of their accounts (“which is analogous to a suspension in traditional work environments”; p. 7). In this case, the AI’s lack of protection abilities as a limitation regarding the AI’s *duty ethics* leads to a control mechanism by humans that ensures their autonomy (*virtue ethics*). There are also examples in which humans use the limitations of the AI to shift blame from themselves to the AI, which demonstrates changes in their *virtue ethics*. For instance, “human team members can hide behind the output of the machine teammate and also use it as a scapegoat” by blaming the AI, even though it is not responsible (Rix and Hess, 2022, 9). They thus “ignore” their own ethical ideology (*virtue ethics*) and delegate all ethical considerations to the AI’s *duty ethics*.

To sum up, how humans (social entity) build perceptions of AI (material entity) depends on the level of influence through the *social subsystem*, the material entity, or the human itself. When considering the *strong* form of *external performativity* in particular, management practices can directly influence humans’ perceptions of AI with respect to possible ethical issues and potential.

Materiality: How interdependencies between objects of the social subsystem and the material entity affect ethical management of human-AI interaction

While discussions on *materiality* are quite diverse, in this work we refer to *materiality* with respect to how the material entity is arranged based on an artifact's physical and non-physical entities and objects. Materiality does not focus on a specific type of human-AI interaction, as it does not concern *sociomaterial entanglement* and instead affects the interaction indirectly. We distinguish between three subforms of *materiality*:

- i) *Physical materiality*: the material entity is arranged based on the (same or other) artifact's physical objects (e.g., documents, hardware, products – representing non-human objects).
- ii) *Non-physical materiality*: the material entity is arranged based on the (same or other) artifact's non-physical objects that can be related either to non-human objects (e.g., processes, machine activities, management practices, guidelines) or human objects (e.g., workforces' norms, leadership values).
- iii) *Digital materiality*: can be subordinated to *non-physical materiality* and states that the material entity is arranged based on the (same or other) artifact's digital objects (e.g., data, software – representing non-human objects)

The following two sections provide examples from the literature review to bolster our suggestion to differentiate between the three subforms to better explain interdependencies between the objects of the *social subsystem* and the material entity.

Concerning *physical materiality*, there are interdependencies between the material entity and the artifact's physical objects. For instance, the AI can be influenced by the user interface as an “important tool for scrutinizing the evidence behind the output” (Someh et al., 2022, p. 156). This is a characteristic of *strong physical materiality* between the AI as the material entity and the interface as (non-human) physical object, given that the material entity is interrelated with the same or other artifact's physical objects. How managers drive AI optimization and intent to use AI as a leadership tool (Watson et al., 2021) represents either *weak* (there are no effects) or *strong* (the material entity affects the same or other artifact's physical objects or vice versa) *physical materiality* between the AI as the material entity and the management as a (human) physical object.

Concerning *non-physical materiality*, there are interdependencies between the material entity and the same or other artifacts' non-physical objects (including digital objects). For instance, data (digital objects) represents a human's discriminating behavior, building on “low-quality input data can lead to biased or unusable outputs”

(Asatiani et al., 2020, p. 270). This can evoke “biased standards” of the AI as material entity, leading to “improper procedures” in the AI’s decision-making (Zhang et al., 2020, pp. 5888–5889). Thus, AI (material entity) depends on data quality (digital object), which is characterized by *strong digital materiality*. This dependency, though, may also go the other way – for instance, when the data output (digital object) can be improved by the AI (material entity) (Klumpp, 2018). These examples highlight that non-physical objects and the material entity can affect each other in a *weak* (there are no effects) or *strong* (the material entity affects other non-physical objects or vice versa) way.

These interdependencies between the objects of the *social subsystem* and the AI as the material entity confirm the importance of differentiating between several forms of *materiality*. In the following, we elaborate on the two forms of the notion *materiality* (*physical materiality* and *non-physical materiality*, also including *digital materiality*) and highlight how they affect the two different ethical perspectives (*virtue* and *duty*) in the management of human-AI interaction.

We begin with *weak* or *strong physical materiality*. The organizational infrastructure, such as technical infrastructure, represents a physical object that can be interpreted from the *duty ethics* perspective, given that “the organizational infrastructure aligns expectations and goals around AI fairness” (Madaio et al., 2020, p. 7). Moreover, discussing guidelines for the design of AI within the workforce may better “address ethical issues both in the initial design and post-launch stage of AI-system” (Wang et al., 2020, p. 4968). Guidelines embedded in the AI through design represent the *duty ethics* perspective of the artifact’s physical objects. Further, empowering ethical guidelines (physical objects) through training as physical objects (comprising the training plan based on *duty ethics*, which might be influenced by the trainer’s *virtue ethics*) (Kelley, 2022, p. 880) can increase the explainability of the AI (Asatiani et al., 2020, p. 274). Consequently, “organizations should offer sufficient training” to make employees “feel capable of understanding the new level of technological complexity” and “democratize the process in such a way that it is actually accessible to everyone” (Rix and Hess, 2022, p. 10). The degree of influence between guidelines and training of the *social subsystem* and the increased transparency of the AI are linked to either *weak* or *strong physical materiality*.

Concerning *weak* or *strong non-physical materiality* in the context of *transparency & accountability*, feedback and explanations (non-physical objects; linked to *strong non-physical materiality*) provided by the workforce as human objects in the *social subsystem* during the use of AI systems (e.g., predicting diagnosis in healthcare) can be used to mitigate ethical issues, for instance, concerning perceived understandability of the AI logic (*duty ethics*) that can occur through this *strong non-physical materiality* (Zhang et al., 2020). In other words, feedback provided by the workforce based on their norms and values (collective *virtue ethics*) influences the AI's logic (*duty ethics*) and can lead to either *strong* or *weak non-physical materiality*. In addition, “discussion through workshops that involve multiple stakeholders” can shed light on the behavior of “Black-Box AI Systems” (Asatiani et al., 2020, p. 259). Discussions among trainees and trainers (*virtue ethics*) are relevant to create “human-consumable explanations,” given that “humans ask different explanation-seeking questions” when interacting with AI depending on their background and the task for which they are using AI (Liao et al., 2020, p. 10). This may lead to the establishment of ethical principles in the organization (*duty ethics*) or increase the ethical understanding of the workforce (collective *virtue ethics*). *Virtue ethics* can be further institutionalized if “managers at all levels are key stakeholders in training intelligent systems so that the machines can complement managers’ knowledge and work habits” (Kolbjørnsrud et al., 2017, p. 5). Hence, to increase transparency and explainability of the AI, there are several mechanisms (representing non-physical objects of the social subsystem) that can affect the *virtue ethics* of human objects and the *duty ethics* of non-human objects on a collective level.

To sum up, organizations might consider the many opportunities to influence AI's *duty ethics* from the *social subsystem* that indirectly contribute to ethical human-AI interaction. Materiality in this case helps us better explain the interdependencies between objects of the *social subsystem* and the AI (material entity).

Practices: How routines of the material entity or emotions of the social entity affect the social subsystem and, thus, the ethical management of human-AI interaction

Practices are created through interaction and, thus, represent an outcome of the notion *inseparability*, wherefore they can be linked to all types of human-AI interaction. *Practices* describe real cases and practices that occur through human-AI interaction in daily work. We distinguish between two subforms of practices:

- i) *Affective practices*: perceived changes of the material entity within sociomaterial entanglement evoke emotions of human object in the social subsystem (e.g., the workforce).
- ii) *Routinized practices*: practices of the material entity within sociomaterial entanglement lead to routinized patterns of non-human objects in the social subsystems (e.g., working operations).

The following two sections provide examples to bolster our suggestion to differentiate between the two subforms of the notion *practices*, given that the differentiation sheds light on how the emotions of individuals as the social entity and routines of the AI as the material entity occur as a result of *sociomaterial entanglement* can affect the *social subsystem*. We also link the subforms of the notion to the types of human-AI interaction.

Beginning with *affective practices*, we determined that in a *strong* form, it can be linked to *substitution* as one type of human-AI interaction. For instance, the authors of a study of Uber suggested that drivers whose driving routes are dictated by algorithms and who receive unexpectedly low ratings may be frustrated because they know this can lead to account deactivation (Pregenzer et al., 2021). This represents *strong affective practices*, meaning that perceived changes of the material entity evoke strong emotions of the social entity in real cases based on the lack of opportunity to intervene in the algorithmic management through the platform (Lee, 2018; Lee et al., 2015). In a *weak* form, the human-AI interaction does not provoke emotions of the social entity in everyday work. *Strong affective practices* with emotions, especially negative ones, occur mainly when certain mechanisms for interventions are substituted through AI.

Regarding *routinized practices*, we observed that this subform of *practices* could often be linked to *augmentation* as one type of human-AI interaction. For instance, being managed by algorithms can change the habits of Uber drivers, who begin to explain to their passengers the meaning of their ratings and try to influence those passengers to leave positive ratings (Pregenzer et al., 2021, p. 7). We refer to this as *strong routinized practices*, meaning that practices of the material entity lead to novel routinized patterns of the social entity in everyday work. In a *weak* form, interacting with AI does not lead to routines among the workforce, such as the non-professionalized use of AI-based digital assistants of single employees in organizations with a low AI maturity (Someh et al., 2022). Hence, *strong routinized practices* occur, especially when it comes to *augmentation*.

These examples demonstrate that there are different ways that *practices* may be created in everyday work based on human emotions or routinized behavior. In the following, we elaborate on the two *strong* forms of the notion *practices* (*strong affective practices* and *strong routinized practices*) and highlight how they affect the two different ethical perspectives (*virtue* and *duty*) of the *social subsystem*. As *weak* forms of this notion demonstrate that interacting with AI does not affect the *social subsystem* over the long term, we do not elaborate on the *weak* form.

We recognized that *strong affective practices* are often linked to *humanity*, as individuals may feel unease in the impersonal AI-based environment created by “the lack of two-way communication” (Acikgoz et al., 2020, p. 14). As a consequence, human objects of the *social subsystems* try to compensate for the AI’s lack of humanity (representing an ethical issue regarding its *duty ethics*) that an individual might perceive during human-AI interaction. Furthermore, social support from other employees as human objects is needed to mitigate that emotional burden. Another example shows that social support can also be a “starting point of worker resistance” (Jiang et al., 2021, p. 412) when “[s]ome drivers try to reassure other drivers by arguing that it is not worth constantly worrying about their ratings” (Jiang et al., 2021, p. 9). Hence, human objects in the *social subsystems* aim to compensate *humanity* for the AI based on their *virtue ethics*.

When analyzing *strong routinized practices*, we identified that changes in habits and processes could be linked primarily to *privacy & (non-)maleficence* since humans adapt their routines toward non-discriminating behavior (*virtue ethics*) based on the principles according to which the AI is programmed (*duty ethics*). For instance, in the context of assigning passengers to Uber drivers, the algorithm’s matching and routing system avoids cherry-picking. This affects the habits of employees, leading to “non-discriminating” behavior of drivers when picking up passengers, since “all rides [are] good for business, in the long run” (Pregenzer et al., 2021, p. 6). However, to follow AI-based principles, it can be helpful to have a “type of checklist that you go through for accuracy,” given that organizations “need a good way of incorporating [fairness] as part of the workflow” (Holstein et al., 2019, p. 8).

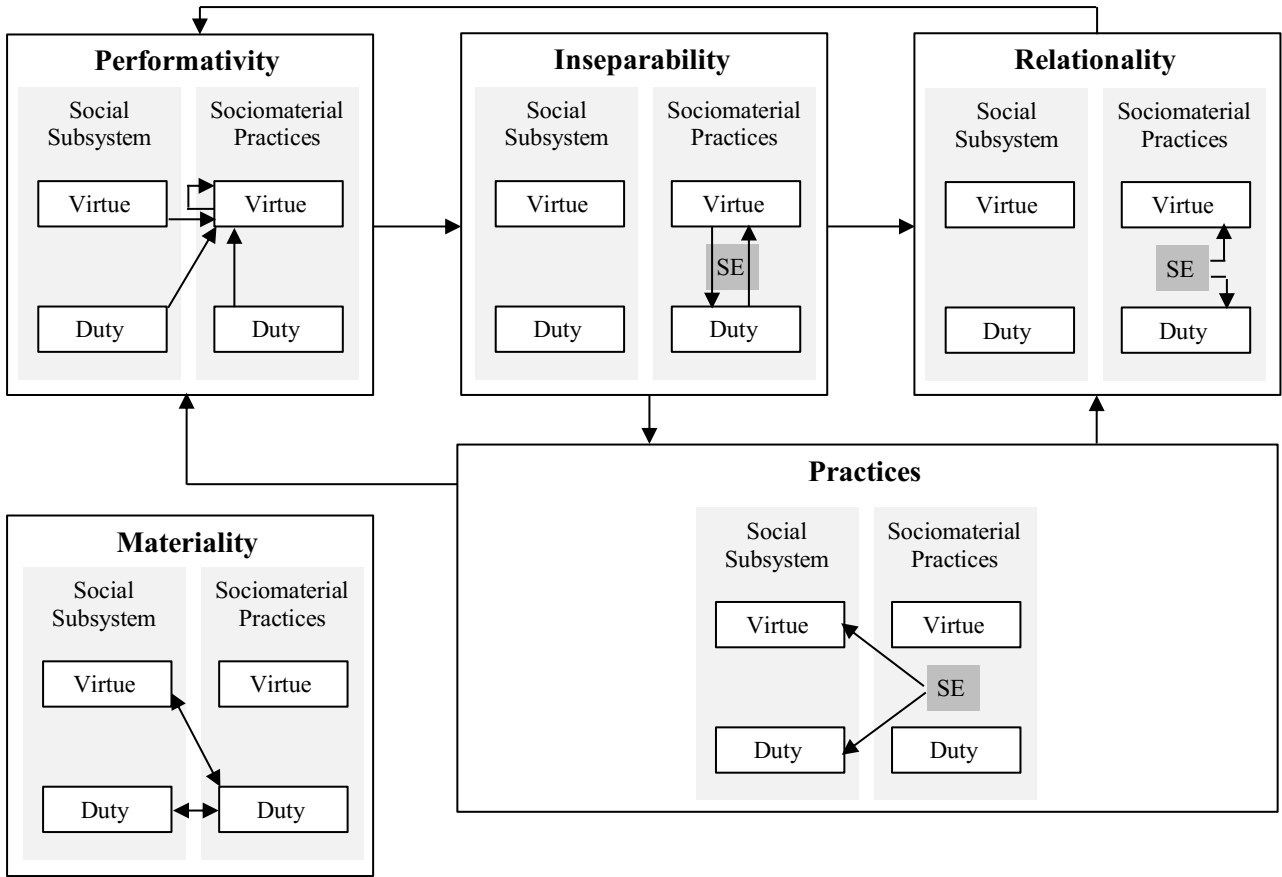
Another example is that “the AI system is designed to pick up information that employees might accidentally have missed, skipped, or omitted when they processed a customer’s loan application,” thus

minimizing mistakes (Strich et al., 2021, p. 316). However, the AI's equal treatment through its non-discrimination mechanism (*duty ethics*) can nonetheless lead to discrimination against "certain groups of people, based on their social status, origin or place of residence" (Mayer et al., 2020, p. 251), since the AI is not aware of situational exceptions. Thus, *routinized practices*, including the non-discriminating ethical ideology of the employee (*virtue ethics*), are required to compensate for the systematic exclusion through the AI-based system (*duty ethics*).

To sum up, several mandatory principles (*duty ethics*) can help increase individuals' ethical behavior (*virtue ethics*). However, there must be some kind of balance since the AI's *duty ethics* can also drive unethical behavior of the individual. There are different ways *practices* can be created in the *social subsystem* based on human emotions or routinized behavior occurring as a result of *sociomaterial entanglement*. Especially when considering *strong practices*, management must be aware of the interdependencies between the ethical intention of an individual's behavior and the *duty ethics* of AI, as both can result in ethical or unethical routines or emotional patterns of the *social subsystem*.

The Sociomaterial Cycles: How the notions of sociomateriality affect ethical management of human-AI interaction through their circular relationship incorporating virtue ethics and duty ethics

An additional result of analyzing the literature through the different subforms of the five notions of sociomateriality is that we determined that the notions act to some degree in a circular relationship within the sociotechnical system. Figure 4 is an overview of these relationships affecting both ethical perspectives (*duty* and *virtue*). This can be used to demonstrate the complexity of human-AI interaction and the upstream and downstream processes surrounding the *sociomaterial entanglement*:



Note: SE = Sociomaterial Entanglement; Virtue = Virtue Ethics; Duty = Duty Ethics

Figure 4: The Sociomaterial Cycles

In the following, we explain the relationships between the notions illustrated in Figure 4. In general, we determined two sociomaterial cycles.

We begin with the first sociomaterial cycle and the notion *performativity*, which demonstrates how limitations and boundaries of the material entity, the social entity itself, and the *social subsystem* influence humans' perceptions of AI (social entity) before interacting with the AI as the material entity (*sociomaterial entanglement*). In doing so, *performativity* can outline how an individual's *virtue ethics* builds on human perceptions. Thus, *performativity* describes interdependencies creating AI perceptions of humans that affect human-AI interaction (*sociomaterial entanglement*) as an antecedent. It influences the moment of entanglement that can be illustrated through the notion *inseparability*.

In comparison, *inseparability* focuses on the moment of "imbrication" as the interaction of the social entity and the material entity during *sociomaterial entanglement*. Within *sociomaterial entanglement*, the *duty ethics*

of the AI as the material entity and the *virtue ethics* of the human as the social entity influence each other through interaction.

As an outcome of this interaction (and thus, as an outcome of *inseparability*), *relationality* concerns change in the characteristics, agency, or existence of the social entity or material entity. Thus, *relationality* demonstrates how the outcome of human-AI interaction can affect humans' *virtue ethics* and AI's *duty ethics*. Given that these changes in *relationality* can also influence the perceptions of humans as the social entities, *relationality* can affect *performativity*, wherefore we observe the first cycle of influences between *performativity*, *inseparability*, and *relationality*.

In addition, *practices* are created through interaction (and thus represent an outcome of the notion of *inseparability*). Since *practices* lead to routinized patterns of non-human objects (incorporating their *duty ethics*) and emotions of human objects (incorporating their *virtue ethics*) that can be investigated in real everyday work environments, they become part of the *social subsystem* and, thus, influence the perceptions of the social entity as a form of *external performativity*. *Practices* can also change the characteristics of the social entity by provoking emotions that lead to adaptations in the behavior of humans, thus representing *relationality of characteristics* of the social entity. These interdependencies between *inseparability*, *performativity*, *relationality*, and *practices* represent another cycle of influence.

Finally, *materiality* moderates these cycles, as it comprises the several objects of the *social subsystem* (comprising *virtue ethics* and *duty ethics*) that are interdependent with the AI as the material entity (comprising *duty ethics*), which is directly involved within all other notions of sociomateriality through the entanglement with the social entity. Hence, it moderates all other notions of sociomateriality, as it mainly represents management guidelines, practices, and strategies to guide the ethical management of human-AI interaction.

By presenting the interrelationships of the notions of sociomateriality in the form of sociomaterial cycles showing how both ethical perspectives (*duty* and *virtue*) are embedded in those notions, we hope to contribute theoretically to increasing the understanding of the ethical management of human-AI interaction.

6. Discussion

Research on ethics in human-AI interaction has long been part of the literature (e.g., Abdelaal, 2021, Boada et al., 2021, Giermindl et al., 2022). This research, though, has tended to focus on *duty ethics* (e.g., Chao, 2019, Hamilton et al., 2021, O'Sullivan et al., 2019) – that is, an individual's or technology's "ethical doing" – without considering the influence of moral assumptions and values of an individual's "ethical being" (*virtue ethics*). To investigate the ethical management of human-AI interaction holistically, we systematically analyzed the literature on this topic in the context of both perspectives (*virtue ethics* and *duty ethics*). We applied both perspectives to the five notions of sociomateriality that reflect the complexity of human-AI interaction organizations aim to manage ethically. Where needed, we introduced subforms of the five notions to allow for a more differentiated and reflective analysis. We concur with the suggestions by Berente et al. (2021) regarding the relevance of applying a sociotechnical perspective to guide ethical considerations in managing human-AI interaction.

6.1. Theoretical Implications: A Research Agenda on Ethical Management of Human-AI Interaction

Our review highlights the most significant contributions that research has made toward understanding the ethical management of human-AI interaction. This section extends these contributions by outlining a future IS research agenda comprising four main avenues. The first avenue explains four effects between *virtue ethics* and *duty ethics* occurring through the arrangement of humans, technologies, and human and non-human objects in organizations. The second highlights the contribution of the sociomaterial notions to study the complexity of human-AI interaction. The third demonstrates the relevance of a more human-centered approach when analyzing the ethical management of human-AI interaction in IS research. The fourth avenue proposes using the "Next-Generation Theorizing" of Burton-Jones et al. (2021, p. 303) to enhance the theorizing part of literature reviews according to state-of-the-art requirements for novel complex phenomena, given the changing nature of human-AI interaction that is introducing complexity to strategic management (Benbya et al., 2020). At the end of our discussion of the first and second avenues below, we provide a table that summarizes opportunities for future research.

Avenue 1: The interdependencies between virtue ethics and duty ethics may shape the ethical management of human-AI interaction

Given that one's ethical understanding and intrinsic motivation to behave ethically (*virtue ethics* of the social entity) can be influenced by ethical obligations (*duty ethics* of the material entity or the *social subsystem*), and vice versa, both perspectives shape the ethical management of human-AI interaction. We found four effects when considering both *virtue ethics* and *duty ethics* in a sociotechnical system: *the responsibility attribution effect*; *the control coherence effect*; *the altruistic behavior effect*; and *the behavior adoption effect*. Details are given below.

The Responsibility Attribution Effect during sociomaterial practices: humans (social entity) shifting responsibility toward AI (material entity) and, thus, curtailing their ethical being (virtue ethics)

The first effect we outline is the *responsibility attribution effect*, which implies, for example, that humans transfer their responsibility for decisions to the AI-based system to avoid unpleasant situations in which they have to justify certain actions or decisions (Bansal et al., 2021, p. 12). In other words, humans (social entities) transfer their agency of being responsible for certain actions to the material entity's agency (*weak material inseparability*). This happens especially with individuals who have higher levels of sociability, given that "individuals who are inherently more social tend to view their technological teammates in more social ways ... [and] are likely to consider the machine as international and agentic." (Jia et al., 2022, p. 394) One striking example of the *responsibility attribution effect* can be seen in the context of a bank where the AI-based loan system takes over decisions to reject or accept loan requests. The bank employee stated, "I think it is quite nice in the case of a loan rejection, because you don't have a long discussion with the customer, the decision is final" (Strich et al., 2021, p. 316).

Other authors refer to an AI acting as a "moral crumple zone taking on responsibility that would have otherwise been assigned to the human" (Hohenstein and Jung, 2020, p. 15). By "feeling that they [decision makers] are no longer responsible for the decisions helps many communicate the decision ... more quickly and easily" (Mayer et al., 2020, p. 246). Without AI-based advice, employees "had to develop their own strategies for communicating rejections" (Mayer et al., 2020, p. 246). When AI makes suggestions, employees tend not to question and judge their decisions in everyday work (*strong routinized practices*), and their "ability to reflect critically on their work" decreases (Mayer et al., 2020, p. 249). Employees "no longer have the opportunity to

develop reflection skills in work-related decisions because the AI system has taken over the entire decision-making process,” so the employees “become indifferent to the outcome of the process.” (Mayer et al., 2020, p. 249) For example, with the growing capabilities of LLM-driven conversational agents such as ChatGPT (Mollick, 2022), user may overtrust their output because they view it as close to perfect. This may lead to users simply accepting and adopting the output generated by the AI without reflecting on it and its consequences critically (Dwivedi et al., 2023). In the long term, their “awareness [may be] severely impaired or even nonexistent” (Mayer et al., 2020, p. 249). This demonstrates how an individual’s values and abilities of critical thinking and judging (*virtue ethics*) can be affected by attributing responsibility for decision-making and actions from the human to an AI. Still, an AI is guided by guidelines (*duty ethics*) that determine how to handle difficult customer situations (Mayer et al., 2021). As a consequence, we can say that transferring responsibility to an AI agency to make guideline-based decisions affects an individual’s ability to judge ethically and critically reflect upon decisions and outcomes (Basu et al., 2021).

The *Control Coherence Effect* between the social subsystem and sociomaterial practices: shifting leadership tasks (*duty ethics*) from the social subsystem to AI (material entity) can make humans (social entity) behave unethically (*virtue ethics*)

We call the second effect of interrelations between *virtue ethics* and *duty ethics* the *control coherence effect* – referring in particular to the notion *material inseparability*. This effect occurs when the AI takes over leadership functions, such as monitoring performance outcomes. Given that the AI as the material entity can perform only certain leadership functions by, for instance, collecting data, it often results in controlling the human rather than monitoring the performance of work processes, leading to control coherence. Humans may try to escape the control mechanism of an AI (*duty ethics*) through loopholes (indicating “ill-defined” or nonexistent control mechanism) or blending behavior since they perceive the tight monitoring as unjustified and a loss of their autonomy (*virtue ethics*) (Pregenzer et al., 2021). By having control over the employee, “computers no longer have a subordinate or supporting role but adopt leadership functions” (Hoeddinghaus et al., 2021, p. 7), which can result in employees “feeling exposed” (Rix and Hess, 2022, p. 9) and perceiving a “reduced level of control” (Asatiani et al., 2020, p. 274). This occurs in particular when we observe an interaction between *weak* and *strong material inseparability*, since loopholes provide possibilities for the social

entity to have workplace practices that are not mediated by AI as the material entity (*weak material inseparability*) (Pregenzer et al., 2021, p. 11). Yet, when there is a very tight control mechanism of the AI, and when humans are highly dependent on the material entity, the AI inherits the leadership functions from the *social subsystem* through its control mechanism and, thus, gains autonomy over the employee (*strong material inseparability*). Even though individuals may perceive automated leadership decisions as more transparent compared to human decisions, the changes in power and authority may result in individuals “perceive[ing] the traditional form of leadership ... as more trustworthy” (Hoeddinghaus et al., 2021, 27).

“Tightening control by introducing even more rules and surveillance may close some loopholes but won’t reduce the desire to resist” (Pregenzer et al., 2021, pp. 11–12). As a result, opaque behavior of human-AI interaction (Mayer et al., 2020; Zhang et al., 2020) or “blending behavior” can occur (Pregenzer et al., 2021, p. 10). For instance, in a study of Uber it was outlined that algorithmic-managed drivers who know the algorithm by which they are managed very well may enter “fake inputs” to get preferred passengers, thus optimizing rating scores as a mechanism to escape the pressure to perform due to constant control (Lee et al., 2015, p. 1611). Hence, we suggest that the higher the performance pressure through tight control mechanisms of an AI and the higher the transparency of those control mechanisms (*duty ethics*), the more likely individuals feel exposed and behave in a non-transparent manner by taking advantage of loopholes to escape control (*virtue ethics*). This demonstrates *augmentation* as one type of human-AI interaction (see section 4), in which humans as the social entity try to exploit the algorithm as the material entity with such loopholes to increase their personal benefits (Lee, 2018; Lee et al., 2015; Zhang et al., 2020). In doing so, the human’s ethical ideology allows for subordinating morally correct behavior to the advantage of exploiting AI’s weakness to escape the control mechanism that they usually allocate to the responsibility of the *social subsystem*. Hence, when AI takes over leadership functions previously done by humans in the *social subsystem*, employees try to escape the control mechanism of the AI’s doing (*duty ethics*) by exploiting AI’s weaknesses during the interaction (representing the human agency’s *virtue ethics*).

The *Altruistic Behavior Effect* during sociomaterial practices: humans (social entity) questioning the ethical doing (duty ethics) of the AI (material entity) based on their ethical being (virtue ethics) can enforce ethical behavior of human-AI interaction

The third effect of interrelationships between *virtue ethics* and *duty ethics* occurs in particular reference to the notions of sociomateriality *material inseparability* and *routinized practices*. We call it the *altruistic behavior effect*, which implies that humans use their knowledge to question decisions made by AI to achieve greater perceived fairness or ensure security – especially for the benefit of others. In this way, humans as the social entity use their ethical understanding to overturn (*weak material inseparability*) or intervene (*strong material inseparability*) decisions made by the AI as the material agency.

While AI is programmed to treat everyone equally and, thus, follows the non-discrimination principle (*duty ethics*), there are some cases in which humans decide differently based on their inherent ethical understanding (*virtue ethics*). In the context of recruitment, algorithms do not have the ability to, for instance, discern good candidates because they lack “human intuition” by making “judgments based [only] on keywords [making them] not capable of evaluating tones of voices or human interaction” (Lee, 2018, 9). Consequently, humans should intervene (Malik et al., 2021) and rely on their *virtue ethics* to evaluate the AI-based decision (Lee et al., 2015; Pregenzer et al., 2021) and, in doing so, “improve [overall] fairness” in the process of human-AI interaction by combining own fairness perceptions with the fairness being programmed in AI (Holstein et al., 2019, p. 9). Hence, we suggest that when the equal treatment of an AI leads to discrimination, humans as social entities can compensate by using their ethical knowledge for sense-making. When the equal treatment of an AI leads to the potential for non-discrimination, the ethical abilities of humans as social entities can be strengthened.

Fostering humanity in the process of human-AI interaction can support overall fairness – even though humans sometimes “outsmart” the AI to balance insufficient moral behavior of the AI having *augmentation* as an underlying type of human-AI interaction. For instance, at the bank previously mentioned, “some loan consultants were misusing the CleverLoan AI system by trying to outsmart it. They justified these actions as either necessary to overcome the restrictions imposed on their field of activity or important for granting a loan to individuals well-known to them” (Mayer et al., 2020, p. 250). Hence, “outsmarting” is about weighing different moral values against each other (e.g., helping others vs. behaving according to transparency standards), whereby the social entity subordinates the ethical principle to act transparently by discarding the AI-based instruction to increase the greater good (thus fostering beneficence).

The *Behavior Adoption Effect* between the material entity and the social subsystem: managers (social subsystem) adopt ethical decisions (duty ethics) of the AI (material entity)

The fourth effect of interrelationships between *virtue ethics* and *duty ethics* refers, in particular, to the notions of sociomateriality *routinized practices*, *relationality of agency*, and *material inseparability*. We call it the *behavior adoption effect*, which implies that humans adopt the behavior of an AI by, for instance, also starting to collect sensitive data for performance evaluation, just as the AI-based performance tracker does (Park et al., 2021). In this way, humans as the social entity adapt their ethical understanding of how to evaluate others to the material entity's ethical agency (*strong routinized practices*). For instance, how the AI is programmed can lead to an ignorance of data privacy concerns in a private situation by "collecting and analyzing ... sensitive information" (Park et al., 2021, p. 8). One interviewee of Park et al. (2021, p. 8) stated, "Let's say you have a stomachache and stay longer in the restroom. You mean the AI will input that data too? I think that's a serious privacy invasion." Another argues, "Don't you think it's too forceful that it could expose even my private feelings that I want to hide from the manager or firm?" Hence, individuals working with AI or being managed through AI may begin to adopt the ethical or unethical behavior of the AI (based on its *duty ethics*) over the long term as they become used to how AI evaluates performance, for instance. In that way, as their ethical understanding (*virtue ethics*) adjusts to the behavior of the AI, individuals may begin to act similarly.

Ethical issues are further reinforced when the programmed data collection rules of the material entity (*duty ethics*) foster unethical behavior of the social entity when using the information for managerial tasks (*virtue ethics*) as *routinized practices*. If managers also begin to collect such sensitive data and use it in managing employees (*strong routinized practices*), they adapt to the misconduct of the AI (*strong material inseparability*). The "organizational anchoring of ethical AI implementation" (Mayer et al., 2021, p. 9) and the "policies on ethical governance considering socially preferable approaches" (Wang et al., 2020, p. 4968) (*duty ethics*) are important mechanisms of the *social subsystem* that can help avoid humans adapting to misconduct in AI and reduce the ethical issue of the *behavior adoption effect*. Including multiple stakeholders with different ethical understandings, such as domain experts, can also help (Asatiani et al., 2020, p. 274). Hence, we propose that the (un)ethical behavior of AI based on its programmed rules to collect data and make decisions can influence

the (un)ethical behavior of managers (and their *virtue ethics* being part of the *social subsystem*) over the long term. Ethical norms and values within the *social subsystem* can moderate those effects.

These effects underline the relevance of understanding how one's inherent ethical ideology is interwoven with the principle-based ethics of the *social subsystem* and the material entity. Hence, future IS research on the ethical management of human-AI interaction should consider the *virtue ethics* perspective more deeply as a way to provide a holistic view of ethics.

Table 1 is an overview of possible research regarding *virtue ethics* in the context of sociomateriality.

Effect	Research Questions
Agency Attribution Effect	• How do humans perceive their own responsibility in AI-based decision-making procedures? • How do humans attribute responsibility to the AI, depending on the nature of the task? • Under which circumstances do individuals shift responsibility toward AI?
Control Coherence Effect	• Under which circumstances does human agency dominate material agency and vice versa? • What are the suppressing effects of AI's control mechanism for the outcomes of human-AI interactions (e.g., trust, appreciation, well-being)? • To what extent can the ethical behavior of humans compensate for the non-ethical behavior of AI during augmentative tasks?
Altruistic Behavior Effect	• What patterns of altruistic behavior exist in human-AI interaction? • What positive and negative consequences occur through altruistic behavior in human-AI interaction? • What can or should organizations do to support positive outcomes of altruistic behavior in human-AI interaction?
Behavior Adoption Effect	• How does AI's ethical behavior affect a manager's ethical ideology? • What might be positive and negative consequences when humans begin to adopt AI's ethical view and behavior?

Table 1: Toward a Research Agenda for Virtue Ethics in Sociotechnical System Theory

Avenue 2: How the nuanced sociomaterial notions contribute to the analysis of human-AI interaction

The second avenue proposes to apply the sociomaterial notions as a theoretical foundation to make sense of the complexity of ethical considerations of human-AI interaction. With it, we demonstrate the appropriateness of using the lens of sociomateriality – that organizations can be described as an arrangement of humans, technologies, and other objects (such as principles and management practices) that are somehow entangled, and that how they are arranged leads to interdependencies and, thus, to different patterns of actions (Cecez-Kecmanovic et al., 2014; Contractor et al., 2011), such as to ethical or non-ethical behavior. We applied different notions of sociomateriality to provide a deeper and more differentiated understanding of the ethical management of human-AI interaction. The sociomaterial notions contribute to explaining how the different ethical perspectives (*virtue* and *duty*) behave in specific human-AI interactions. Similar to Jones (2014), we identified inconsistencies when analyzing the ethical management of human-AI interaction through the two perspectives of *weak* and *strong* sociomateriality (e.g., the same human-AI interaction can represent *weak* and *strong* *relationality* depending on whether we reflect the perspective of the material or social entity).

The several ways in which AI interacts with humans are too complex to draw pragmatic differentiations, given the “variety of AI use cases” (Figuerola-Armijos et al., 2022, p. 12) and various tasks for which AI is used that “impact ethical perceptions” (Figuerola-Armijos et al., 2022, p. 1). For example, in the context of autonomous driving, humans are often unaware of AI; a human’s cognitive limitations may lead to not recognizing that AI might be part of the car (Hinsen et al., 2022). Instead, humans see the artifact or technology (e.g., application) in which the AI is embedded, such as the vehicle that is parking automatically, thanks to the driving assistant (Hinsen et al., 2022). This example demonstrates *weak material performativity*, since the boundaries of the social entity can lead to perceiving AI as the material entity as nonexistent. Therefore, no perceptions regarding AI’s involvement arise when interacting with the technology in which the AI is embedded. But *strong physical materiality* is also illustrated in the example, as the AI changes the vehicle by giving it the agency to behave autonomously.

This example also addresses different ethical aspects. It shows the need for greater transparency in the design of AI-based applications (*duty ethics*) to raise humans' awareness of AI as an important step to making AI more usable in organizations (Someh et al., 2022). In the context of algorithmic-managed leadership decisions in particular, making the use of the algorithm to enhance leaders' decision-making transparent for employees increases their awareness regarding AI involvement, which can lead to increased perceived fairness and trustworthiness of individuals (Hoeddinghaus et al., 2021). This may also result in a shift from *strong* to *weak social performativity*, a shift that can affect an individual's cognitive ability to understand how AI influences the AI-based application as an artifact, given the complexity of the underlying methods and algorithms that can lead to cognitive overload for humans. Hence, training is required to increase an individual's knowledge (Hinsen et al., 2022; Mayer et al., 2020), which points to the *strong physical materiality* of the training since it would not exist without the human-AI interaction.

This example demonstrates that a simple distinction between *weak* and *strong* sociomateriality would not fully explain the underlying complexity of human-AI interaction – especially when “people will have to ... understand decisions and consequences from mixed AI/human worlds” (Coombs et al., 2021, p. 4). Therefore, we suggest that the systematic investigation of human-AI interaction through the sociomaterial notions helps increase our understanding of the ethical management of human-AI interaction. By further specifying the single notions, we hope to contribute significantly to the current theoretical discourse. For the same complexity reasons, differentiating solely between different types of human-AI interaction (such as *substitution*, *augmentation*, *assemblage*) as archetypes is not sufficient. As argued in section 4, we have to be aware of how the different tasks are embedded in the process of human-AI interaction. The types of human-AI interaction consider only the moment of sociomaterial entanglement in a specific task (and thus lack considerations of antecedents and consequences of this entanglement), whereas the sociomaterial notions enable a more differentiated analysis of the entire human-AI interaction process (involving the antecedents and consequences of this entanglement), as demonstrated through the sociomaterial cycles.

Table 2 is an overview on future research regarding how the notions of sociomateriality contribute to our understanding of human-AI interaction.

Notion	Research Questions
Materiality	• What is the role of AI ethics in different governance structures of organizations (e.g., codes of conduct, Corporate Social Responsibility guidelines)? (<i>physical materiality</i>) • To what extent does AI ethics affect the ethical values (<i>virtue ethics</i>) on a collective level (e.g., organizational culture)? (<i>non-physical materiality</i>) • To what extent does AI ethics affect the ethics of other technological objects (e.g., design principles of machine ethics)? (<i>digital materiality</i>)
Inseparability	• How can we align AI values (<i>duty ethics</i>) with human values (<i>virtue ethics</i>)? (<i>social inseparability</i>) • How can we measure the alignment between human values and AI values? (<i>material inseparability</i>)
Relationality	• How can we measure positive outcomes (e.g., satisfaction, self-esteem, well-being) arising through changed existence, characteristics, or agency of individuals as a result of human-AI interaction? (<i>relationality of individuals' characteristics and agency</i>) • To what extent are positive or negative changes in the existence, characteristics, or agency of individuals a result of human-AI interaction? (<i>relationality of individuals' characteristics and agency</i>)
Performativity	• How can training influence human agency? (<i>external performativity</i>) • How does the transparency of AI affect human agency? (<i>material performativity</i>) • How do humans build perceptions of AI? (<i>social performativity</i>)
Practices	• How do emotions arising from human-AI interaction affect an individual's <i>virtue ethics</i>? (<i>affective practices</i>) • How does <i>virtue ethics</i> of individuals differ depending on how they are used for interacting with AI? (<i>routinized practices</i>)

Table 2: Toward a Research Agenda for the Notions of Sociomateriality

Avenue 3: How and why a more human-centered approach contributes to the analysis of human-AI interaction

In this avenue, we strive for a more human-centered approach by providing an overview of current research streams concerning human-AI interaction within different disciplines. We demonstrate how our research objective is embedded in the different research fields.

The computer science literature often lacks a human-centered approach. This became clear given how many articles from this discipline we had to exclude because they focus on the design of AI-based technologies, such as articles that explore how to increase fairness and explainability of the AI (e.g., explainable AI (XAI) studies) but often do not explicitly consider humans as the social entity and how their associations and behavior might also affect fairness and explainability of human-AI interaction. When it comes to ethical considerations in particular, a human-centered approach is vital since humans should not be designed out of the technology (Coombs et al., 2021; Ehsan and Riedl, 2020; Liao et al., 2020). For instance, in the context of clinical AI-assisted decision-making, “human-consumable explanations” are crucial to increase trust in the system’s output; otherwise, humans could be overwhelmed by the AI explanations (Liao et al., 2020, p. 10). Given that “AI will have a degree of [randomness] and may not be 100% explainable,” explainability should be seen as a social process of conversation and interaction between humans and AI to create “tailor[ed] explanations ‘for people with different backgrounds’” (Liao et al., 2020, p. 7). Thus, with respect to the strategic, ethical management of human-AI interaction, “[t]he employee should remain the central element in the workplace, and perceptions of and experiences with AI are fundamental for the process of introducing it to the workplace” (Mirbabaie et al., 2022, p. 87). It is critical, therefore, to ensure a “human-centered” approach for AI given that it “directly impact[s] people’s lives and well-being” (Bankins et al., 2022, p. 40).

This is especially relevant given that because of the “non-human and unemotional nature of the AI” (Bankins et al., 2022, p. 15), humans have doubts about their own value, leading to diminished self-esteem. For instance, “the lack of two-way communication in an AI-based interview may signal to the applicant that they will be treated impersonally as an employee (e.g., as ‘just a number’ or a ‘cog in the machine’), if the company chooses not to take the time to have a personal interview” (Acikgoz et al., 2020, p. 412). When AI interacts with

humans, the latter can have thoughts such as: “How can we be evaluated by a robot? We are humans and should be treated as valuable individuals” or “It feels like ... becoming a part of a gigantic machine” (Park et al., 2021, p. 7). Furthermore, humans sometimes express that their human-AI interactions are “weird, strange, and alien” (Park et al., 2021, p. 7) or that when AI behaves very much like humans, it gives them an “uneasy, creepy feeling” (Rajaobelina et al., 2021, p. 2351). Hence, as algorithms “lack human intuition” (Lee, 2018, p. 9), ethical management needs to preserve key attributes of humanity in human-machine interaction (Coombs et al., 2021) through the *social subsystem*, such as providing emotional support (Kelley, 2022; Park et al., 2021) or feedback loops (Sauppé and Mutlu, 2015). AI’s inability to act in a human-intelligent way, such as by being socially and emotionally intelligent, affects the outcomes of human-AI interaction. By integrating mechanism of the *social subsystem* that compensates for the AI’s inhumanity, we have a more human-centered approach that contributes to the analysis of human-AI interaction.

The overwhelming majority of the articles we reviewed offer answers to our first subquestion (SQ1), which focuses on how AI interacts with humans. In contrast, far fewer articles provide insights into how humans interact with AI (representing SQ2). There are also articles that focus on the human part (especially in the organizational and management sciences), but they often lack a deeper analysis of ethical management since they are most concerned with investigating AI-related trust and AI attitudes. However, there are interdependencies – for instance, regarding the transfer of ethical knowledge between humans and AI. While the AI’s ethical knowledge is expressed in rules and code (e.g., *duty ethics*), the human’s ethical knowledge is gained over years of experience and is not easy to express explicitly (e.g., intuition based on *virtue ethics*). Although AI also learns from experience, it lacks “the capabilities to handle situations that are not documented” (Malik et al., 2021, p. 343). Consequently, humans should intervene (Malik et al., 2021) and use their experience to evaluate AI (Lee et al., 2015; Pregonzer et al., 2021) and manage their interactions ethically. Considering only one perspective (either humans interacting with AI or AI interacting with humans) results in ignoring such interdependencies. Hence, we must integrate both perspectives when investigating ethics in human-AI interaction to answer questions such as how to develop ethical AI based on collective social processes of conversation and interaction between humans and AI.

By looking at the intersection, we determined that this is an emerging field of research, especially given that most of the articles in our final sample have been published since 2020. The amount of qualitative work in our final sample further demonstrates that this is an up-and-coming research area.

Avenue 4: How Next-Generation Theorizing contributes to the theorizing part of theoretical literature reviews

Our work further highlights the contributions of using the “Next-Generation Theorizing” of Burton-Jones et al. (2021, p. 303) to enhance the theorizing part of a literature review. More specifically, the *extend strategy* of “Next-Generation Theorizing” enabled us to employ state-of-the-art requirements in the analysis of the changing nature of human-AI interaction as a complex phenomenon that has introduced additional complexity in strategic management (Benbya et al., 2020). We not only consolidated the insights of existing research but also explained the relationships among the constructs in the conceptual framework. In this way, we extend theory on ethics through the lens of sociomateriality to increase our understanding of the ethical management of human-AI interaction.

Based on the four shifts in views of theorizing of Burton-Jones et al. (2021), we derived possible steps that enhance the theorizing part of our review and correspond to the steps for transparency in literature reviews of Templier and Paré (2018). One shift toward “Next-Generation Theorizing” concerns seeing theory as “a lens” to better understand the complexity in this phenomenon and using it as a “starting assumption” (Burton-Jones et al., 2021, p. 306). To address this shift, we embedded sociomateriality directly into the problem formulation as the first step for literature reviews, following Templier and Paré (2018), since the lens of sociomateriality helps in mapping the complexity of the phenomenon. We further refined sociomateriality by using its notions and deriving subforms of each notion to provide a more differentiated analysis in the sixth step of Templier and Paré (2018). In addition, we conducted the first (problem formulation) and second (literature search) step with a holistic view of ethics by involving interdisciplinary research from several domains in the problem formulation and database as well as our keyword search strategy. However, it should be mentioned that the manner in which we consider ethics also represents a theoretical lens that can, depending on which ethical philosophy some refer

to, lead to other forms of theories. Still, it would be interesting to investigate what other ethical lenses are used in IS research on human-AI interaction.

Another shift toward “Next-Generation Theorizing” treats theory as “intellectual conversation” and exhorts researchers to focus on the fact that some old theoretical contributions remain relevant while others may change “dramatically” (Burton-Jones et al., 2021, p. 306). We take this shift into account in several steps of Templier and Paré (2018). For instance, we did not restrict the literature search to a certain time range or to one specific research domain. We also conducted “intellectual conversations” in the form of several rounds of discussion when analyzing data and interpreting results, thus adapting our framework.

The shift from “theory as purely desk work” to “theory as reflective engagement” addresses how researchers learn “with and from practitioners to theorize important issues” (Burton-Jones et al., 2021, p. 306). Our analysis is based mainly on articles that report and analyze field research and thus treat human-AI interaction as a changing phenomenon in the real world. We aim to reflect practical cases systematically with the theoretical lens of the notions of sociomateriality and sociotechnical system theory. Thus, we try to foster “reflective engagement” by learning through discussions of the different practical cases while theorizing. This could be extended by embedding reflective methods of empirical approaches into the development of theory to involve different ethical perspectives of ethicists, philosophers, managers, and so on. Future research should investigate how to embed reflective methods (e.g., interviews) in the development of theory.

Finally, the shift “theory as performative” addresses the “need to appreciate how the sociomaterial phenomena we are studying is itself partly a product of theorizing” (Burton-Jones et al., 2021, p. 306). Especially when establishing our framework as the main part of the analysis (sixth step of Templier and Paré, 2018), we recognized patterns to follow the call of Jones (2014) for a more differentiated analysis of sociomateriality (see Avenue 2).

6.2. Practical Recommendations: Strategic Management of Ethical Human-AI Interaction

Our review further highlights the relevance of strategic management to fostering ethical human-AI interaction in organizational contexts. In this section, we extend these practical implications by outlining

recommendations for three main avenues for future strategies and practices on the ethical management of human-AI interaction. In doing so, we heed the calls of Vial (2019) and of Alsheibani et al. (2020) to enhance our understanding of ethics with respect to the multilevel impact of digital transformation as a way to provide recommendations for leaders in the strategic management of AI.

Our fifth avenue concerns the relevance of bringing ethics to a strategic level in organizations to manage human-AI interaction. Our sixth avenue aims at making employees more literate in their interactions with AI by fostering their (ethical) knowledge for equality in human-AI interaction. Our seventh avenue suggests integrating different stakeholder views, striving for tailored measures, and fostering interdisciplinary research.

Avenue 5: Bringing ethics to a strategic level in managing human-AI interaction in organizations

Ethics is often discussed at either an *operational level* by integrating ethical guidelines into *operative processes*, such as through crash courses on ethics in technological practice projects (Wang et al., 2020) or by implementing ethical checklists (Kelley, 2022; Madaio et al., 2020); or at a *functional level* by providing training and feedbacks in *tactical processes*, such as through regular feedbacks to alert about possible problems and systematic errors (Mayer et al., 2020). Our review highlights the relevance of integrating ethical considerations into corporate and business strategies. Ethics with respect to human-AI interaction requires a greater strategic perspective; as such, ethical checklists and/or guidelines must be internalized and supported by the organizational culture and integrated into the organization's goals and priorities (Grønsund and Aanestad, 2020; Madaio et al., 2020). In addition, a more “value-sensitive design” – with participatory elements being part of the practice and strategies of an organization – is required to develop a holistic understanding of “who” the human is (Ehsan and Riedl, 2020, p. 16). This also contributes to reducing perceived creepiness when interacting with AI-based technologies by integrating human elements, such as touch points (Rajaobelina et al., 2021).

Following the arguments of Kelley (2022), cultural differences put ethical management of human-AI interaction on the business level. Business strategies with respect to policies for AI principles may vary across different countries and may not depend on an overarching corporate strategy. To put ethics on a strategic level, organizations must leverage resources (Madaio et al., 2020) to establish an organizational culture that aims to foster reflection and critical thinking (Mayer et al., 2020). Ethics becoming a part of the business strategy and

organizational culture leads to a more objective form of *virtue ethics* at an organizational level, one that is aligned with tailored ethical measures of the AI design based on *duty ethics*. Hence, discussing ethics at a strategic level in organizations will contribute positively to the ethical management of human-AI interaction.

Avenue 6: Making the workforce more literate in human-AI interaction

Our review highlights the relevance of fostering the AI literacy of employees who interact with AI. For instance, Polyviou and Zamani (2022) argue for upskilling the working population with AI skills and enhancing education curricula on AI to foster digital literacy on AI (which is also known as AI literacy). In general, AI literacy can be understood as the skills, capabilities, and competencies that enable individuals to interact with AI effectively (Long and Magerko, 2020). Research on AI literacy highlights the relevance of ethics to successful interaction with AI (e.g., Heyder and Posegga, 2021; Ng et al., 2021), which our review confirms. For instance, a lack of experience can make employees feel as if they are not entitled to direct control over the algorithm and hence do not desire it, which can have ethical consequences (Lee et al., 2015). Furthermore, individuals judge the fairness of algorithmic versus human-made decisions based on their opinion of whether they assume the human or the AI to be more qualified (Lee, 2018). Thus, employees need experiences that can help them achieve a more objective assessment rather than leaving them to make decisions based on gut feelings. Employees require a “sufficient level of metaknowledge” when delegating tasks to AI (Fuegener et al., 2021, p. 680).

A basic understanding of AI and essential information regarding the ethical thinking of AI are also important elements to increase the awareness of ethical AI use in an organization (Mayer et al., 2021, p. 8). Mandatory or voluntary workshops and training events, in particular, can “give employees the knowledge they need to make informed ethical decisions about AI” (Mayer et al., 2021, p. 10). In addition, it is important to install a “culture of learning and experimentation” (Grønsund and Aanestad, 2020, p. 10) to drive “employee’s skill and capability renewal” (Watson et al., 2021, p. 471) as well as “adaptive capabilities as humans become involved in new triadic human-AI relationships” (Cranefield and Doyle, 2022, p. 1). While, for instance, “[a]pplications like ChatGPT can be used either as a companion or tutor, to support for example self-regulated learning,” they can also lead to individuals minimizing their effort, and thus reduces their independent thinking skills (Dwivedi

et al., 2023, p. 9). Hence, the transformative character of AI-based technologies can strengthen but also limit the development of an individual's capabilities. In addition, AI applications "may not be able to produce accurate responses to inputs outside of its training data, or that are highly specific or niche" (Pavlik, 2023, p. 90). Therefore, responsible use of recent AI-based technologies requires sensitivity to the AI's lack of creativity and contextual understanding, the biases in its training data, its limited understanding, its limited ability to personalize instructions, and its dependency on data (Baidoo-Anu and Owusu Ansah, 2023).

Education can help to overcome the "risk of an intelligence divide between humans and machines" (Gupta et al., 2021, p. 12) or "artificial divide" – a serious risk for the strategic management of human-AI – by raising the capability level of employees so they can cooperate successfully with the AI application (Klump and Zijm, 2019, p. 270). This is linked to the literature on the AI divide that describes inequalities in the access, usage, and outcomes of human-AI interaction (Carter et al., 2020). Training can strengthen ethical understanding from the bottom up (*virtue ethics*) in addition to top-down approaches such as guidelines and norms (*duty ethics*) (Jiang et al., 2021; Madaio et al., 2020).

Adapting employee skills also requires "new employment schemas," such as new forms of reward systems, to avoid "inequalities and ... social fissures" (Coombs et al., 2021, p. 5). In addition to training to upskill workers (Bankins et al., 2022), it is also important to hire employees with the required skillset (Grønsund and Aanestad, 2020). Hence, fostering AI literacy in the workforce can be understood as a prerequisite to the ethical management of human-AI interaction in organizations.

Avenue 7: Integrating different stakeholder views and striving for tailored measures

Our review outlines that ethics is always subject to interpretation and thus stresses the need for an interdisciplinary "ethics team" to interrogate external guidelines (*duty ethics*) with their expertise (*virtue ethics*) (Grønsund and Aanestad, 2020; Mayer et al., 2020; Wang et al., 2020). Moreover, we wish to highlight "the importance of discussing the 'hard decisions' as a group, which may suggest an AI ethics committee may be better than a single AI ethics officer. Regardless, some form of ethics office(r) is important for the effective adoption of AIPs [AI principles]" (Kelley, 2022, p. 881). Emerging AI-based technologies such as ChatGPT offer unique opportunities to incorporate collective ethical considerations in future human-AI interaction

outcomes through feedback options based on user input (Radford et al., 2018). Such democratic features that employ the wisdom of the crowd can strengthen the AI model's *duty ethics* based on "collective deliberation and decisions" (Etzioni and Etzioni, 2017, p. 413) represented in user feedback and, as a consequence of further use, strengthen the *virtue ethics* of the user. However, we support the suggestion that ethical measures "would need to differ for different teams or organizations to reflect their organizational culture, goals, and priorities" (Madaio et al., 2020, p. 10). This would mean the involvement of "managers from different levels and geographies in initial experiments with AI" (Kolbjørnsrud et al., 2017, p. 6). Therefore, industry associations, managers, and policymakers should get involved when talking about responsible AI (Gupta et al., 2021). In addition to managers, it is also important that data scientists and "various participants with detailed knowledge on specific segments ... join the meetings and contribute their interpretation" (Grønsund and Aanestad, 2020, p. 9). Hence, discussions among different stakeholder groups are relevant "to learn from one another's experiences" (Holstein et al., 2019, p. 8). This form of "cultivating ... new expertise" contributes to a greater explainability of the AI, which also increases the value of AI for the organization (Someh et al., 2022, p. 153).

The diversity of roles, sectors, and application areas of AI demands that ethical guidelines be adaptable to different teams and organizations to "reflect their organizational culture, goals, and priorities" (Madaio et al., 2020, p. 10). Ethics, and *virtue ethics* in particular, is partly subjective, and so organizations must "tailor their AI adoption strategies to local and organizational conditions." Since "all technology is local [i]t's about the employee at the desk or in the cubicle and how they use it" (Kolbjørnsrud et al., 2017, p. 5). Hence, to foster ethical management of human-AI interaction, organizations have to establish an ethical ecosystem, activating multi-stakeholder collaboration when governing AI strategies and implementations (Polyviou and Zamani, 2022), as "collective deliberation and decisions" can strengthen *virtue ethics* on a workforce level (Etzioni and Etzioni, 2017, p. 413).

Where do we go from here? Ethical management of human-AI interaction considering recent advances in generative AI

The public and academic discourse on AI-based systems has recently been refueled by advances in technologies based on generative AI (Pavlik, 2023). More specifically, LLMs now serve as the foundation for

a variety of services that are accessible to a broad audience within and outside of organizations and that can address a plethora of tasks revolving around processing and generating textual information (Pavlik, 2023). The release of Open AI's LLMs, especially GPT3 and GPT4, has set new standards because of their capability to interpret natural language prompts and generate textual responses that are almost indistinguishable from human writing (Pavlik, 2023). The subsequent integration of these models into ChatGPT, OpenAI's conversational agent, has offered a glimpse into the potential of this branch of generative AI and has served as a role model and foundation for various commercial applications that are likely to have a major impact on knowledge work and human-AI interaction (Baidoo-Anu and Owusu Ansah, 2023; Dwivedi et al., 2023). As a result, this novel generation of generative AI is now widely accessible to a wide range of users (Baidoo-Anu and Owusu Ansah, 2023).

However, the rapid spread of these technologies has amplified some of their inherent issues and fueled the current debate on ethical considerations related to their design and use (Greene et al., forthcoming). Due to the universal applicability of these technologies, it is important to understand how to manage the novel forms of human-AI hybrids from an ethical standpoint, both in theory and practice. In this light, our work can serve as a starting point to guide and structure future research on the ethical management of human-AI hybrids.

In Avenue 1, we discuss the interdependencies between *virtue ethics* and *duty ethics* and how they may shape the ethical management of human-AI interaction. AI-based technologies, including prominent examples such as ChatGPT, are subject to multiple sources of ethical principles. Thus, it is important to consider what happens if they are misaligned or even contradict each other. For instance, *duty ethics*, in the form of organizational norms and guidelines as well as external norms and regulations, set boundaries for developers and guide the development of the system. *Virtue ethics* of individual developers and those of users providing feedback by using and training the system may not always be consistent with the system's *duty ethics*. In the case of ChatGPT, ethical alignment issues become amplified due to the system's widespread use, its integration with various other services, and the public attention it receives (Dwivedi et al., 2023). The interplay between various ethical principles at different stages of developing, training, moderating, and using such technologies remains, for the most part, unexplored. With the growing adoption and success of these proprietary systems, it

will be vital to improve our understanding of the ethical influences that shape their behavior to avoid unintended consequences of ethically misaligned models.

This pertains further to Avenues 5 and 7, which address the need to bring ethics to a strategic level of managing human-AI interaction. As Dwivedi et al. (2023, p. 13) stated, “If we want to have any control over this disruption, we need digital leadership and a conversation about acceptable and ethical behavior.” Thus, improving our understanding of related phenomena requires research on the strategic management of human-AI interaction, which in turn requires the consideration of stakeholders on all governance layers, such as at a collective (e.g., society, industries, organizations) and individual level (e.g., users). Research on this subject can benefit from a Next-Generation Theorizing approach, which we suggest for analyzing human-AI interaction in times of rapid innovation (see Avenue 4).

In Avenue 2 and Avenue 3, we discuss how the nuanced sociomaterial notions contribute to analyzing human-AI interaction and why a more human-centered approach contributes to the analysis of human-AI interaction. Both avenues offer potential insights into the novel forms of human-AI hybrids that result from the increasing adoption and maturity of generative AI. Due to their ability to interpret natural language prompts, process large amounts of information, and generate compelling textual responses to tasks provided by their users, systems based on generative AI can take on various roles in human-AI hybrid teams: simple ones such as summarizing, editing, and processing textual information; and more complex roles as (co-)producers of content, innovations, and source code (Dwivedi et al., 2023). Understanding the broad spectrum of roles from simple to complex and the consequences for human-AI interaction, as well as their ethical implications, is vital for grasping the technology’s impact on workplace practices. Our framework offers a starting point for future research in this direction: the nuanced sociomaterial notions provide a structure for further investigating the ethical management of various types of human-AI hybrids (Avenue 2) while promoting a human-centered approach (Avenue 3).

Avenue 6 calls for making the workforce more literate in human-AI interaction. We observe that it is of the utmost importance to consider the context in which generative AI is applied to figure out which skills and capabilities are required to achieve certain kinds of outcomes that are ethically justifiable. Given that we know

little about the skills required to use AI technologies effectively (Long and Magerko, 2020), critically reflect upon their outcomes, and train and develop corresponding “skills, resources, and capabilities to handle generative AI [and] examining biases” (Dwivedi et al., 2023, p. 3), this avenue is particularly promising to advance research on the ethical management of human-AI interaction.

To summarize, we hope that our framework enables the analysis of complex ethical issues that arise with the progress of generative AI, which is leading to novel forms of human-AI interaction. We hope further that our avenues may contribute by integrating a holistic view into the current debate surrounding the use of generative AI applications.

7. Conclusion and Limitations

Our *theoretical review* of the ethical management of human-AI interaction highlights and synthesizes an interdisciplinary and novel body of literature that contributes to our understanding of how we can ethically manage AI interacting with humans (SQ1, concerning how AI’s ethical doing based on *duty ethics* might affect human’s ethical doing and being) and humans interacting with AI (SQ2, concerning how human’s ethical being as *virtue ethics* and doing might affect the AI’s ethical doing). Against the backdrop of the current progress of generative AI in gaining increased autonomy (Dwivedi et al., 2023; Mollick, 2022), we hope to provide valuable suggestions with respect to the ethical considerations that arise when managing the changing nature of human-AI interactions. First, we suggest that a more differentiated and systematic analysis, applying the notions of sociomateriality, contributes to explaining the complexity of the different types of human-AI interaction and further clarifies the interdependencies between the different constructs in sociotechnical systems. Second, our work highlights the importance of increased awareness concerning the interplay between *virtue ethics* and *duty ethics* in the context of managing human-AI interaction. We aimed to make this interplay more tangible with the sociomaterial cycles.

With respect to practical implications, our review outlines the importance of bringing ethics to a strategic level in the organization, making the workforce more literate to deal with AI in an ethical way, and integrating different stakeholder views to achieve a holistic view of ethics. Overall, we sincerely hope our findings prove

valuable both for future research and for practitioners by enriching the understanding of the ethical management of human-AI interaction.

Our work does have several limitations. First, while we aimed to provide systematicity and transparency in the review process by following the steps of Templier and Paré (2018), the main part of our sixth step (analysis and discussion) was built upon intensive discussion rounds only among the three coders. Following the “Next-Generation Theorizing” of Burton-Jones et al. (2021), in particular, future research could enrich discussion rounds by involving practitioners. Second, our review is restricted to articles that exhibit a certain quality (e.g., by fulfilling specific rankings and peer-reviewed criteria). To include most recent research, future researchers could be less strict in their exclusion and inclusion criteria. Third, our work builds upon the specific perspective of sociomateriality, and thus our results depend on that lens. While this is a limitation, we nevertheless find this lens a valuable tool for investigating the ethical management of human-AI interaction in more depth. Future research that follows a different approach may provide valuable insights into our research objective. Fourth, the objective of understanding ethical management of human-AI interaction called for an approach that favored breadth over depth. Even though we demonstrate relationships of the main constructs, we focused on providing an overall holistic picture. Future research on this topic may zoom in on specific relationships that have emerged from the analysis.

Finally, our *theoretical review* highlights the relevance of rethinking the way we manage human-AI interaction in organizations ethically. To cope with the changing nature of human-AI interaction in terms of increased performance, autonomy, acceptance, and quality, organizations must place ethics on a strategic level. To account for the interdependencies between the inherent ethical values of individuals and the principles or methods AI is designed on, research must consider ethics by reflecting on both perspectives: humans interacting with AI and AI interacting with humans. We hope our framework on the ethical management of human-AI interaction provides valuable guidance for future research and helps inform organizations on how to foster ethics in the strategic management of human-AI interaction. Especially in light of the current debates on generative AI (Mollick, 2022) currently surrounding GPT4 with respect to the unsupervised training of models, close cooperation between users and AI is essential to enable the weighing of different ethics of maxima.

To close out our research, we prompted ChatGPT to consider these very issues in brief, using the prompt “Act as a researcher, create a concluding statement regarding the ethical management of human-AI interaction.” Its response was as follows: “*Ensuring ethical management of human-AI interaction is critical for creating a future where humans and machines can work together in harmony. It is our responsibility as researchers to not only develop AI technologies that are reliable and effective but also to consider the ethical implications of their use and ensure that they align with our shared values of fairness, transparency, and accountability.*”

The increased capabilities and implementation of AI-based technologies redefine how we work and how we collaborate, and we sincerely hope to provide valuable contributions to better understand the up-and-coming phenomenon.

Appendices

Appendix A: Notions of Weak and Strong Sociomateriality

In table A.1, we adopt and extend toward our research context the *weak* and *strong* notions of sociomateriality as defined by Jones (2014).

Notion: Inseparability	
Inseparability describes the multiple ways social and material entities can be interwoven (Jones, 2014). It focuses on the process of interaction between the social and the material entity and the role each entity plays for the other during the interaction (limited to the imbrication during the process of fulfilling a specific task or goal). The notion suggests that neither entity can function separately. Although Orlikowski and Scott (2008, p. 434) suggest an “inherent inseparability” of the social and material entity, inseparability is not restricted to a specific ontology, as the interrelationship of both entities can vary between <i>strong</i> and <i>weak</i> (Jones, 2014).	
Strong inseparability describes a “mutual constitution” (Jones, 2014, p. 920). The material and social entities are one unit or, rather, a symbiosis, given that certain tasks or functions do not work without the other entity.	Weak inseparability describes a “mutual interdependency” (Jones, 2014, p. 920). The material and social entities operate as separate units in a certain task or function, and they are able to work without the other entity as well in other functions or tasks.
Notion: Relationality	
<i>Relationality</i> describes how sociomaterial entities (social and material entities) are “being produced through relation” (Jones, 2014, 898). Compared to inseparability, <i>relationality</i> focuses on the outcome resulting from sociomaterial entanglement. It describes how the entities and their agency, as well as their characteristics, may change because of sociomaterial entanglement. Following Thompson (2012),	

<i>relationality</i> can have different perspectives and varies through sociomaterial entanglement since organizations are not “static entities with determinate[d] properties” (Jones, 2014, p. 898).	
Strong relationality describes that “forms, attributes, and capabilities of the entities emerge only through inter-penetration” (Jones, 2014, p. 920). Without sociomaterial entanglement, there are no changes in the agency or the characteristics of the material or social entity, and no novel entities are created.	Weak relationality describes that “forms, attributes, and capabilities of the entities may preexist in any relation” (Jones, 2014, p. 920). Hence, the characteristics of the social or material entity and their agency exist independent of sociomaterial entanglement with the other entity.
Notion: Performativity	
<i>Performativity</i> describes the way how a social entity creates and describes its “reality” (especially linked to the material entity) based on several factors that limit their perceptions (Jones, 2014, p. 899). Different perspectives unfold, limiting rational description. Such boundaries are not “pre-given or fixed, but enacted in practices” (Orlikowski and Scott, 2008, p. 462).	
Strong performativity describes the “enactment of boundaries and relations” (Jones, 2014, p. 920), which means that the existence of the material entity and/or its practices depend on the perception of the social entity. Hence, explanations and limitations of the social entity create the material entity and/or its practices.	Weak performativity describes that material entities and practices of material entities exist without being recognized by social entities (“non-human agency”); (Jones, 2014, p. 920). Hence, the reality of the social entity (humans) does not create the material entity, as it exists independent of the perceptions of the social entity.
Notion: Materiality	
<i>Materiality</i> describes how a material entity (tangible or intangible) is arranged based on an artifact’s physical or non-physical objects (Faulkner and Runde, 2011; Jones, 2014; Leonardi, 2012). Hence, <i>materiality</i> describes how one material entity is arranged around other non-human objects, such as other technologies, principles, or guidelines, of the same artifact or of other artifacts. <i>Materiality</i> also comprises human objects of organizational artifacts, such as culture, language, leadership, or the management board, given that <i>materiality</i> goes beyond the artifact’s physical entity (Barad, 2003; Orlikowski, 2010).	
Strong materiality describes the “materialization of the phenomena” (Jones, 2014, p. 920). The investigated material entity creates new objects within its artifact or changes existing objects across place and time.	Weak materiality describes the “persistence of the arrangement of material entities across place and time” (Jones, 2014, p. 920). The investigated material entity does not affect the objects of other artifacts – that is, not changing them across place and time (or making them disappear) – but it may affect the objects of those artifacts.
Notion: Practices	
<i>Practices</i> refer to several forms of mental or physical activities, emotions, knowledge, and so on (Reckwitz, 2002) that can be investigated in everyday work. It refers how perceived changes of the material entity within sociomaterial entanglement affect human and non-human objects of the social subsystem in a real environment (Jones, 2014, p. 914).	
Strong practices are about “embodied, materially, mediated arrays of human activity” (Jones, 2014, p. 920). The material entity mediates physical labor by changing the way of working of the social entity that goes beyond activities by altering human working habits (from an internal standpoint, that is, behavior-related changes).	Weak practices focus on “activities and processes” (Jones, 2014, p. 920). The material entity only mediates activities and processes of the entanglement (from an external standpoint, that is, process-related changes).

Table A.1: Notions of Weak and Strong Sociomateriality, based on Jones (2014)**Appendix B: Quality Assessment Criteria Used for Framework Development**

The assessment criteria we used to develop a framework with desirable quality are based on Schwarz et al. (2007). The way we correspond to those criteria being relevant to our work is given in table B.1.

Quality criteria for frameworks aggregated and adapted from Schwarz et al. (2007, pp. 42–43)	Author Considerations
Set up clear guidelines and be aware of the limitations and boundaries of the framework.	- We focused on an <i>integrating type</i> of conceptual framework following MacInnis (2011), wherefore the framework does not focus on stating issues or drawing assumptions at this stage in the research process. Only after the second stage the framework was updated and enhanced. - We did not claim for comprehensiveness in the <i>conceptual stage</i> (in comparison to the <i>review stage</i>).
Identify relevant research streams and collect different theories.	- We searched in several databases incorporating literature from various research disciplines (see Appendix C) and used relevance-filter options of the databases (if available). - We clarified our main constructs: “human-AI interaction,” “ethics,” and “strategic management” as our main research streams and restricted the search process to those main constructs (see Appendix M-O). - We involved different perspectives of <i>weak</i> and <i>strong</i> sociomateriality, and, thus, considered relational as well as substantial ontology (Cecez-Kecmanovic et al., 2014).
Involve all relevant concepts and focus on them.	- We used the references that were most often cited and provided theoretical foundations to identify relevant concepts (Weber, 2012) following a snowball sampling strategy based on backward searches. - We ignored articles that do not use theory for the same purpose as we aim to use in our work, especially those that does not consider strategic or management thoughts (Larsen et al., 2019).
Explain ideas on the relationship between the relevant concepts.	- We explained the interdependencies of our main constructs through the integration of our main constructs in the sociotechnical system of Leonardi (2012). - We further explained interdependencies within the subconstructs of our main constructs following the suggestions of MacKenzie (2003).
Organize specific concepts already studied in several research streams and derive assumptions related to the structure of the framework.	We organized subconstructs and constructs of the framework by embedding them into the sociotechnical system of Leonardi (2012) in section 4.
Use the framework as a lens to make sense of the synthesized literature.	We used the arrangement of our main constructs in the sociotechnical system to explain their interdependencies regarding the ethical management of human-AI interaction. This lens is used in the <i>conceptual stage</i> as well as in the <i>review stage</i> of our research as it is appropriate to think about ethical considerations occurring in the “non-neutral ... performativity of the social and technological entanglement” (Cecez-Kecmanovic et al., 2014, p. 826)
Establish a framework that assists to further theory-development.	- We used the framework as an architecture that guides the second stage of our research process following the idea of integrating conceptualizations of MacInnis (2011). - By following the <i>extend strategy</i> for theory development of Burton-Jones et al. (2021) we used our framework that reflects well-established theories on ethics and sociomateriality (see section 2) and combined it with the relevant subconstructs that helps us to explain the new phenomenon concerning the changing nature of human-AI interaction (see section 4).

Table B.1: Quality Criteria for Framework Development**Appendix C: Leading Journals and Conferences**

Leading conferences and journals of the Information Systems (IS) and the Management and Organizational Science research domain are categorized according to the different rankings given in table C.1:

Leading Outlet Title	Rank	Database
IS-Basket (referring to the Association for Information Systems, 2011): A*/A-Ranking		
MIS Quarterly (Management Information System)	A*	ACM (1977-2019), AIS (1977-present), EBSCO (1977-present), JSTOR (1977-2015), WoS (1980-present)
ISR (Information Systems Research)	A*	ACM (1990-2020), EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)
JMIS (Journal of Management Information Systems)	A	ACM (1984-2012), EBSCO (1984-present), JSTOR (1984-2013), WoS (1999-present)
JSIS (Journal of Strategic Information Systems)	A	ACM (1992-2018), EBSCO (2005-2010), SciDir (1991-present), WoS (1995-present)
JIT (Journal of Information Technology)	A	EBSCO (2004-2020), WoS (1993-present)
JAIS (Journal of the Association of Information Systems)	A	ACM (2000), AIS (2000-present), EBSCO (2003-present), WoS (2006-present)
ISJ (Information Systems Journal)	A	ACM (2012-2017), EBSCO (1998-present), WoS (1991-present)
EJIS (European Journal of Information Systems)	A	ACM (1991-2008), EBSCO (1997-present), WoS (1993-present)
TOP IS-Journals (referring to the Lowry et al., 2013): A+/A-Ranking		
MIS Quarterly (Management Information System)	A+	ACM (1977-2019), AIS (1977-present), EBSCO (1977-present), JSTOR (1977-2015), WoS (1980-present)
ISR (Information Systems Research)	A+	ACM (1990-2020), EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)
JMIS (Journal of Management Information Systems)	A+	ACM (1984-2012), EBSCO (1984-present), JSTOR (1984-2013), WoS (1999-present)
EJIS (European Journal of Information Systems)	A	ACM (1991-2008), EBSCO (1997-present), WoS (1993-present)
ISJ (Information Systems Journal)	A	ACM (2012-2017), EBSCO (1998-present), WoS (1991-present)
JAIS (Journal of the Association of Information Systems)	A	ACM (2000-present), AIS (2000-present), EBSCO (2003-present), WoS (2006-present)
JSIS (Journal of Strategic Information Systems)	A	ACM (1992-2018), EBSCO (2005-2010), SciDir (1991-present), WoS (1995-present)
Leading IS-Conferences (referring to the VHB-JOURQUAL3): A-Ranking		
ICIS (International Conference on Information Systems)	A	AIS (1980-2020), EBSCO (1981-2006)
Leading IS-Conferences (referring to the ERA Australian Research Council, Category: Information Systems): A-Ranking		
CSCW (AMC Conference on Computer Supported Cooperative Work)	A	ACM (1986-2020)
CIKM (ACM International Conference on Information and Knowledge Management)	A	ACM (1993-2020), EBSCO (2008-2020), WoS (2012-2019)
SIGIR (ACM International Conference on Research and Development in Information Retrieval)	A	ACM (1971-present), EBSCO (2000)
UIST (ACM Symposium on User Interface Software and Technology)	A	ACM (1988-2020), WoS (2012-2019)

OCIS (AoM Organizational Communication and Information Systems)	A	Not included, check via manual search
ACIS (Australasian Conference on Information Systems)	A	AIS (2001-2020), WoS (2000-present)
HCI (British Computer Society Conference on Human-Computer Interaction)	A	ACM (2007-2015)
CSCL (Computer Supported Collaborative Learning)	A	ACM (1995-2009)
CIDR (Conference on Innovation Data Systems Research)	A	Not included, check via manual search
SCC (IEEE International Conference on Service Computing)	A	ACM (2012-present), IEEE (2004-2020)
ICWS (IEEE International Conference on Web Services)	A	ACM (2004-2005), IEEE (2004-2020), WoS (2004-2019)
ISWC (IEEE International Symposium on Wearable Computing)	A	ACM (2013-2020), IEEE (1997-2012)
VL/HCC (IEEE Symposium on Visual Languages and Human-Centric Computing (was VL))	A	ACM (2009-present), IEEE (2008-2020), WoS (2011-2020)
BPM (International Conference in Business Process Management)	A	Not included, check via manual search
CaiSE (International Conference on Advanced Information Systems Engineering)	A	ACM (2006-present), WoS (1995-2019)
CoopIS (International Conference on Cooperative Information Systems)	A	IEEE (1996-1999), WoS (1997-2019)
DESIST (International Conference on Design Science Research in Information Systems and Technology)	A	ACM (2009-present)
VHB 2015 (referring to the HARZING-list, Category: General & Strategy): A+/A-Ranking		
Academy of Management Journal	A+	EBSCO (1963-present), JSTOR (1963-2015), WoS (1980-present)
Academy of Management Review	A+	EBSCO (1976-present), JSTOR (1976-2015), WoS (1983-present)
Administrative Science Quarterly	A+	EBSCO (1956-present), JSTOR (1956-2017), WoS (1980-present)
Academy of Management Annals	A	EBSCO (2007-present), WoS (2007-present)
Journal of Management	A	EBSCO (1975-present), SciDir (1993-2004), WoS (1983-present)
Journal of Management Studies	A	EBSCO (1964-present), WoS (1980-present)
Organizational Research Methods	A	EBSCO (1999-present), WoS (1998-present)
Strategic Entrepreneurship Journal	A	EBSCO (2007-present), WoS (2007-present)
Strategic Management Journal	A	EBSCO (1980-present), JSTOR (1980-2015), WoS (1980-present)
VHB 2015 (referring to the -list, Category: Management Information Systems, Knowledge Management): A+/A-Ranking		
Information Systems Research	A+	ACM (1990-2020), EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)
MIS Quarterly	A+	ACM (1977-2019), AIS (1977-present), EBSCO (1977-present), JSTOR (1977-2015), WoS (1980-present)
European Journal of Information Systems	A	ACM (1991-2008), EBSCO (1997-present), WoS (1993-present)

Health Care Management Science	A	EBSCO (1998-present), WoS (2007-present)
IIE Transactions	A	EBSCO (1996-2017), WoS (1982-2016)
Information Systems Journal	A	ACM (2012-2017), EBSCO (1998-present), WoS (1991-present)
Journal of Information Technology	A	EBSCO (2004-2020), WoS (1993-present)
Journal of Management Information Systems	A	ACM (1984-2012), EBSCO (1984-present), JSTOR (1984-2013), WoS (1999-present)
Journal of Strategic Information Systems	A	ACM (1992-2018), EBSCO (2005-2010), SciDir (1991-present), WoS (1995-present)
Journal of the Association for Information Systems (AIS)	A	AIS (2000-present), EBSCO (2003-present), WoS (2006-present)
SIAM Journal on Computing	A	ACM (1984-2017), EBSCO (1981-present), WoS (1980-present)
VHB 2015 (referring to the HARZING-list, Category: Operation Research, Management Sciences, Production & Operations Management): A+/A-Ranking		
Management Science	A+	ACM (1954-2018), EBSCO (1954-present), JSTOR (1954-2015), WoS (1980-present)
Operations Research	A+	ACM (1956-2020), EBSCO (1956-present), JSTOR (1956-2015), WoS (1980-present)
European Journal of Operational Research	A	EBSCO (1977-present), SciDir (1977-present), WoS (1980-present)
INFORMS Journal on Computing	A	ACM (1996-2020), EBSCO (1996-present), WoS (1999-present)
Journal of Operations Management	A	EBSCO (1980-present), SciDir (1980-2018), WoS (1998-present)
Journal of Scheduling	A	ACM (2003-2019), WoS (2001-present)
Journal of the Royal Statistical Society, Series A	A	EBSCO (1974-present), JSTOR (1948-1987)
Journal of the Royal Statistical Society, Series B	A	EBSCO (1998-present), JSTOR (1948-1997)
Manufacturing and Service Operations Management	A	EBSCO (1999-present)
Mathematical Programming	A	ACM (1971-2019), EBSCO (1980-present), WoS (1980-present)
Mathematics of Operations Research	A	ACM (1976-2018), EBSCO (1976-present), WoS (1980-present), JSTOR (1976-2015)
OR Spectrum	A	ACM (1979-2019)
Production and Operations Management	A	EBSCO (2002-present), WoS (1999-present)
Transportation Science	A	EBSCO (1967-present), JSTOR (1967-2015), WoS (1980-present)
VHB 2015 (referring to the HARZING-list, Category: Organization Behavior/Studies, Human Resource Management, Industrial Relations): A+/A-Ranking		
Organization Science	A+	EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)
Journal of Organizational Behavior	A	EBSCO (1988-present), JSTOR (1988-2015), WoS (1988-present)
Leadership Quarterly	A	EBSCO (1996-present), SciDir (1990-present), WoS (1994-present)
Organization Studies	A	EBSCO (1980-present), WoS (1981-present)
Organizational Behavior and Human Decision Processes	A	EBSCO (1966-present), SciDir (1985-present), WoS (1985-present)

Personnel Psychology	A	EBSCO (1948-present), WoS (1980-present)
ABDC 2019 (referring to the HARZING-list, Category: General & Strategy): A*/A-Ranking		
Academy of Management Annals	A*	EBSCO (2007-present), WoS (2007-present)
Academy of Management Journal	A*	EBSCO (1963-present), JSTOR (1963-2015), WoS (1980-present)
Academy of Management Review	A*	EBSCO (1976-present), JSTOR (1976-2015), WoS (1983-present)
Administrative Science Quarterly	A*	EBSCO (1956-present), JSTOR (1956-2017), WoS (1980-present)
Journal of Management	A*	EBSCO (1975-present), SciDir (1993-2004), WoS (1983-present)
Journal of Management Studies	A*	EBSCO (1964-present), WoS (1980-present)
Organizational Research Methods	A*	EBSCO (1999-present), WoS (1998-present)
Strategic Management Journal	A*	EBSCO (1980-present), JSTOR (1980-2015), WoS (1980-present)
Academy of Management Discoveries	A	EBSCO (2015-present)
Academy of Management Perspectives	A	EBSCO (1987-present), WoS (2006-present)
Australian Journal of Management	A	EBSCO 1976-present), WoS (2007-present)
British Journal of Management	A	EBSCO (1990-present), WoS (2000-present)
Business & Society	A	EBSCO (1988-present), WoS (2008-present)
Business Strategy & the Environment	A	EBSCO (1995-present), WoS (2009-present)
California Management Review	A	EBSCO (1958-present), WoS (1980-present)
Global Strategy Journal	A	EBSCO (2012-present), WoS (2011-present)
Group Decision and Negotiation	A	EBSCO (1997-present), WoS (1995-present)
Harvard Business Review	A	ACM (1985-1998), EBSCO (1922-present), WoS (1980-present)
International Journal of Conflict Management	A	EBSCO (2006-present), WoS (1994-present)
International Journal of Management Reviews	A	EBSCO (1999-present), WoS (2001-present)
Journal of Forecasting	A	EBSCO (1982-present), WoS (1982-present)
Journal of Management Inquiry	A	EBSCO (1992-present), WoS (1995-present)
Journal of Sport Management	A	EBSCO (1987-present), WoS (1993-present)
Labour History (Australia)	A	EBSCO (1980-2002), WoS (2007-2020)
Long Range Planning	A	EBSCO (1968-present), SciDir (2002-present), WoS (1980-present)
Management and Organization Review	A	EBSCO (2005-present), WoS (2008-present)
Management Learning	A	EBSCO (1970-present), WoS (1994-present)
MIT Sloan Management Review	A	EBSCO (1960-present), WoS (2001-2020)
Organization & Environment	A	EBSCO (1998-present), JSTOR (1997-2017), WoS (2001-present)

Sport Management Review	A	EBSCO (2009-present), WoS (2011-present)
Strategic Entrepreneurship Journal	A	EBSCO (2007-present), WoS (2007-present)
Strategic Organization	A	EBSCO (2003-present), JSTOR (2003-2017), WoS (2007-present)
Strategy Science	A	SciDir (2021-present)
Theory, Culture & Society	A	EBSCO (1982-present), WoS (1991-present)
ABDC 2019 (referring to the HARZING-list, Category: Management Information Systems, Knowledge Management): A*/A-Ranking		
ACM Transactions on Computer Human Interaction	A*	ACM (1994-present), EBSCO (2005-present), WoS (2007-present)
Decision Support Systems	A*	ACM (1989-2018), EBSCO (1995-present), WoS (1991-present), SciDir (1985-present)
European Journal of Information Systems	A*	ACM (1991-2008), EBSCO (1997-present), WoS (1993-present)
Information & Management	A*	ACM (1979-2020), EBSCO (1978-present), WoS (1983-present), SciDir (1977-present)
Information Systems Journal	A*	ACM (2012-2017), EBSCO (1998-present), WoS (1991-present),
Information Systems Research	A*	ACM (1990-2020), EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)
Insurance, Mathematics & Economics	A*	EBSCO (1985-present), SciDir (1982-present), WoS (1982-present)
International Journal of Information Management	A*	ACM (1989-2020), EBSCO (1984-present), SciDir (1986-present), WoS (1986-present)
Journal of Information Technology	A*	EBSCO (2004-2020), WoS (1993-present)
Journal of Management Information Systems	A*	ACM (1984-2012), EBSCO (1984-present), JSTOR (1984-2013), WoS (1999-present)
Journal of Strategic Information Systems	A*	ACM (1992-2018), EBSCO (2005-2010), SciDir (1991-present), WoS (1995-present)
Journal of the Am. Soc. For Information Science and Technology	A*	ACM (2001-2013), EBSCO (2001-2013), WoS (2001-2013)
Journal of the Association for Information Systems (AIS)	A*	AIS (2000-present), EBSCO (2003-present), WoS (2006-present)
MIS Quarterly	A*	ACM (1977-2019), AIS (1977-present), EBSCO (1977-present), JSTOR (1977-2015), WoS (1980-present)
AIS Transactions on Human-Computer Interaction	A	AIS (2009-present)
Annales of Applied Probability	A	EBSCO (2020-present)
Behaviour and Information Technology	A	EBSCO (1982-present), WoS (1985-present)
Communications of the ACM	A	ACM (1958-present), EBSCO (1965-present), WoS (1980-present)
Communications of the AIS	A	ACM (1999-2000), AIS (1999-present)
Computers in Human Behaviour	A	SciDir (2021-present)
Data & Knowledge Engineering	A	ACM (1989-2017), EBSCO (1987-2013), WoS (1994-present), SciDir (1985-present)
Electronic Commerce Research	A	ACM (2000-2019), EBSCO (2002-present), WoS (2008-present)

Electronic Markets	A	ACM (2007-2008), WoS (2009-present)
Human Computer Interaction	A	EBSCO (1985-present), WoS (1994-present)
IBM Systems Journal	A	ACM (1962-2008), EBSCO (1966-2008), WoS (1980-2008)
IIE Transactions	A	EBSCO (1996-2017), WoS (1982-2016)
Information and Software Technology	A	ACM (1989-2018), EBSCO (1977-present), SciDir (1987-present), WoS (1987-present)
Information Society (The)	A	ACM (1986-2017), EBSCO (1981-present), WoS (1997-present)
Information Systems Frontiers	A	ACM (1999-2019), EBSCO (2002-present), WoS (2001-present)
Information Technology and People	A	EBSCO (1998-present), WoS (2009-present)
International Journal of Electronic Commerce	A	ACM (1996-2012), EBSCO (2000-present), JSTOR (1996-2013), WoS (2000-present),
Journal of Computer Information Systems	A	EBSCO (2000-present), WoS (1993-present)
Journal of Database Management	A	ACM (1994-2018), EBSCO (1993-present), WoS (2004-present)
Journal of Enterprise Information Management	A	EBSCO (2004-present), WoS (2015-present)
Journal of Global Information Management	A	ACM (1993-present), EBSCO (1993-present), WoS (2005-present)
Journal of Knowledge Management	A	EBSCO (1998-present), WoS (2009-present)
Journal of Organizational Computing and Electronic Commerce	A	ACM (1996), EBSCO (1991-present), WoS (1999-present)
Journal of Quality Technology	A	ACM (1986-present), EBSCO (1969-present), WoS (1980-present)
Knowledge Based Systems	A	EBSCO (1980-2020), WoS (1991-present)
MIS Quarterly Executive	A	AIS (2002-present), EBSCO (2006-present), WoS (2008-present)
Scandinavian Journal of IS	A	ACM (1991-2006), AIS (1989-2020), EBSCO (2009-present)
ABDC 2019 (referring to the HARZING-list, Category: Operation Research, Management Sciences, Production & Operations Management): A*/A-Ranking		
Annals of Probability	A*	JSTOR (1973-2017), WoS (1980-present)
Annals of Statistics	A*	EBSCO (2003-2020), JSTOR (1973-2017), WoS (1980-present)
Biometrics	A*	EBSCO (1946-present), JSTOR (1947-2015), WoS (1980-present)
Biometrika	A*	EBSCO (1902-present), JSTOR (1901-2015), WoS (1980-present)
Decision Sciences	A*	EBSCO (1970-present), WoS (1984-present)
European Journal of Operational Research	A*	EBSCO (1977-present), SciDir (1977-present), WoS (1980-present)
Journal of Operations Management	A*	EBSCO (1980-present), SciDir (1980-2018), WoS (1998-present)
Journal of the Royal Statistical Society, Series B	A*	EBSCO (1998-present), JSTOR (1948-1997)
Management Science	A*	ACM (1954-2018), EBSCO (1954-present), JSTOR (1954-2015), WoS (1980-present)
Manufacturing and Service Operations Management	A*	EBSCO (1999-present)

Production and Operations Management	A*	EBSCO (2002-present), WoS (1999-present)
Transportation Research Part A: Policy & Practice	A*	EBSCO (1993-present), SciDir (1976-present), WoS (1980-present)
Transportation Research Part B: Methodological	A*	EBSCO (1999-present), SciDir (1979-present), WoS (1980-present)
Transportation Research Part C: Emerging Technologies	A*	EBSCO (2004-present), SciDir (1993-present), WoS (1995-present)
Transportation Research Part E: Logistics and Transportation Review	A*	EBSCO (1997-present), WoS (1997-present)
Advances in Applied Probability	A	WoS (1980-present), JSTOR (1969-2017)
Annals of Operations Research	A	ACM (1988-1993), EBSCO (1984-present), WoS (1991-present)
Applied Statistics: Journal of the Royal Statistical Society Series C	A	EBSCO (1965-present), JSTOR (1952-2016), WoS (1982-1997)
Computers & Industrial Engineering	A	ACM (1986-2020), EBSCO (1985-present), SciDir (1976-present), WoS (1980-present)
Computers & Operations Research	A	EBSCO (1989-present), SciDir (1974-present), WoS (1980-present)
IEEE Transactions on Engineering Management	A	EBSCO (1966-present), WoS (1980-present)
IEEE Transactions on Intelligent Transportation Systems	A	EBSCO (2012-2019), WoS (2000-present)
Industrial Management and Data Systems	A	EBSCO (1981-present), WoS (1994-present)
International Journal of Logistics Management	A	EBSCO (1990-present), WoS (2008-present)
International Journal of Operations & Production Management	A	EBSCO (1980-present), WoS (1994-present)
International Journal of Physical Distribution & Logistics Management	A	EBSCO (1994-present), WoS (2008-present)
International Journal of Production Economics	A	EBSCO (1991-present), SciDir (1991-present), WoS (1991-present)
International Journal of Production Research	A	EBSCO (1961-present), WoS (1980-present)
International Journal of Project Management	A	EBSCO (1997-present), SciDir (1983-present), WoS (2009-present)
Journal of Business Logistics	A	EBSCO (1978-present), WoS (2008-present)
Journal of Multivariate Analysis	A	ACM (1986-2020), EBSCO (1976-present), SciDir (1971-present), WoS (1980-present)
Journal of Optimization Theory & Applications	A	ACM (1967-2019), EBSCO (1975-present), WoS (1980-present)
Journal of Productivity Analysis	A	EBSCO (1989-present), JSTOR (1989-2015), WoS (1994-present)
Journal of Purchasing and Supply Management	A	EBSCO (2003-present), SciDir (2003-present), WoS (2009-present)
Journal of Scheduling	A	ACM (2003-2019), WoS (2001-present)
Journal of Service Management	A	EBSCO (2009-present), WoS (2009-present)
Journal of Supply Chain Management	A	EBSCO (1999-present), WoS (2008-present)
Journal of the Operational Research Society	A	EBSCO (1978-present), JSTOR (1970-1977), WoS (1980-present)
Journal of the Royal Statistical Society, Series A	A	EBSCO (1974-present), JSTOR (1948-1987)

Journal of Transport Geography	A	EBSCO (2002-present), SciDir (1993-present), WoS (2006-present)
Mathematics of Operations Research	A	ACM (1976-2018), EBSCO (1976-present), JSTOR (1976-2015), WoS (1980-present)
OMEGA – International Journal of Management Science	A	SciDir (1973-present), WoS (1980-present)
Operations Research	A	ACM (1956-2020), EBSCO (1956-present), JSTOR (1956-2015), WoS (1980-present)
Operations Research Letters	A	ACM (1981-2020), EBSCO (2002-present), SciDir (1981-present), WoS (1983-present)
Production Planning & Control	A	EBSCO (1986-present), WoS (1994-present)
Reliability Engineering & System Safety	A	EBSCO (2008-2009), SciDir (1988-present), WoS (1981-present)
Research Technology Management	A	EBSCO (1993-present), JSTOR (1988-2013), WoS (1988-present)
Safety Science	A	EBSCO (1984-present), SciDir (1991-present), WoS (1991-present)
Supply Chain Management: An International Journal	A	WoS (2003-present)
Technometrics	A	ACM (1986-2000), EBSCO (1959-2019), JSTOR (1959-2015), WoS (1980-present)
Theory and Decision	A	WoS (1980-present)
Transport Policy	A	EBSCO (1994-present), SciDir (1993-present), WoS (2005-present)
Transport Reviews	A	EBSCO (1982-present), WoS (1984-present)
Transportation	A	EBSCO (1995-present), WoS (1980-present)
Transportation Research Part D: Transport & Environment	A	EBSCO (1996-present), Sci Dir (1997-present), WoS (1996-present)
Transportation Science	A	EBSCO (1967-present), JSTOR (1967-2015), WoS (1980-present)
ABDC 2019 (referring to the HARZING-list, Category: Organization Behavior/Studies, Human Resource Management, Industrial Relations): A*/A-Ranking		
British Journal of Industrial Relations	A*	EBSCO (1966-present), WoS (1980-present)
Human Relations	A*	EBSCO (1960-present), WoS (1980-present)
Human Resource Management (US)	A*	EBSCO (1972-present), WoS (1985-present)
Industrial and Labor Relations Review	A*	EBSCO (1947-present), JSTOR (1947-2017), WoS (2012-present)
Industrial Relations	A*	EBSCO (1932-present), WoS (1980-present)
Journal of Conflict Resolution	A*	EBSCO (1965-present), JSTOR (1957-2017), WoS (1980-present)
Journal of Human Resources	A*	EBSCO (1966-present), JSTOR (1966-2015), WoS (1980-present)
Journal of Organizational Behavior	A*	EBSCO (1988-present), JSTOR (1988-2015), WoS (1988-present)
Journal of Vocational Behavior	A*	EBSCO (1971-present), SciDir (1971-present), WoS (1980-present)
Leadership Quarterly	A*	EBSCO (1996-present), SciDir (1990-present), WoS (1994-present)
Organization Science	A*	EBSCO (1990-present), JSTOR (1990-2015), WoS (1990-present)

Organization Studies	A*	EBSCO (1980-present), WoS (1981-present)
Organizational Behavior and Human Decision Processes	A*	EBSCO (1966-present), SciDir (1985-present), WoS (1985-present)
Personnel Psychology	A*	EBSCO (1948-present), WoS (1980-present)
Australian Journal of Labour Law	A	Not included, check via manual search
Business Ethics Quarterly	A	EBSCO (1991-present), JSTOR (1991-2015), WoS (2001-present)
European Journal of Industrial Relations	A	EBSCO (1999-present), WoS (1996-present)
European Journal of Work and Organizational Psychology	A	EBSCO (1996-present), WoS (2005-present)
Family Business Review	A	EBSCO (2005-present), WoS (2005-present)
Gender, Work and Organisation	A	EBSCO (1994-present)
Group and Organization Management	A	EBSCO (1976-present), WoS (1992-present)
Human Performance	A	EBSCO (1988-present), WoS (1994-present)
Human Resource Management Journal (UK)	A	EBSCO (1997-present), WoS (2009-present)
Human Resource Management Review	A	EBSCO (1991-present), SciDir (1991-present), WoS (1994-present)
Industrial Relations Journal	A	EBSCO (1970-present)
International Journal of Human Resource Management	A	EBSCO (1990-present), WoS (2000-present)
International Journal of Industrial Organization	A	EBSCO (1983-present), SciDir (1983-present), WoS (1987-present)
International Journal of Intercultural Relations	A	EBSCO (1977-present), SciDir (1977-present), WoS (1983-present)
International Journal of Manpower	A	EBSCO (1980-present), WoS (1994-present)
International Journal of Selection and Assessment	A	EBSCO (1993-present), WoS (1995-present)
International Negotiation: A Journal of Theory and Practice	A	EBSCO (1996-2017)
Journal of Business and Psychology	A	EBSCO (1986-present), JSTOR (1986-2015), WoS (1994-present)
Journal of Business Ethics *	A	EBSCO (1982-present), JSTOR (1982-2017), WoS (1982-present)
Journal of Industrial Relations	A	EBSCO (1959-present), WoS (2007-present)
Journal of Occupational & Organizational Psychology	A	EBSCO (1991-present), WoS (1992-present)
Labor History	A	EBSCO (1960-present), WoS (1980-present)
New Technology, Work and Employment	A	EBSCO (1997-present), WoS (1994-present)
Organization	A	EBSCO (1994-present), WoS (1995-present)
Organizational Dynamics	A	EBSCO (1972-present), SciDir (1972-present), WoS (1980-present)
Organizational Psychology Review	A	EBSCO (2011-present), WoS (2011-present)

Personnel Review	A	EBSCO (1993-present), WoS (1980-present)
Psychology of Women Quarterly	A	EBSCO (1976-present), WoS (1980-present)
Research in Organizational Behavior	A	EBSCO (1979-2019), SciDir (2000-2019), WoS (1982-2018)
Small Group Research	A	EBSCO (1973-present), WoS (1990-present)
Stress and Health	A	EBSCO (1983-present), WoS (2001-present)
Work & Occupations	A	EBSCO (1974-present), WoS (1982-present)
Work & Stress	A	EBSCO (1987-present), WoS (1987-present)
Work, Employment & Society	A	EBSCO (1987-present), JSTOR (1987-2017), WoS (1991-present)
<i>WoS = Web of Science, SciDir = Science Direct</i>		

Table C.1: Leading Journals and Conferences of the IS Domain and Management and Organization

Science Domain

Leading conferences and journals of ethics domain were categorized according to different rankings given in table C.2. Since ethical journals and conferences have a lower or no ranking compared to the IS and Management and Organizational Science domain, the rank differs, although they are the leading conferences and journals in the ethics domain.

Leading Outlet Title	Rank	Database
VHB 2015 (referring to the VHB-JOURQUAL3, Category: Ethics): B/C-Ranking		
Journal of Business Ethics	B	EBSCO (1982-present), WoS (1982-present)
Business Ethics Quarterly (BEQ)	B	EBSCO (2001-present), WoS (2001-present)
Journal of Business, Economics & Ethics	C	Not included, check via manual search
Business Ethics: A European Review	C	EBSCO (1998-present), WoS (2008-2020)
Ethical conferences (referring to the ERA Australian Research Council, Category: Ethics): listed as relevant, but not ranked		
Conference on AI, Ethics and Society (AIES)	-	SciDir (2020-present), WoS (2020-present)
Ethics and Information Technology	-	WoS (2001-present)
<i>WoS = Web of Science, SciDir = Science Direct</i>		

Table C.2: Leading Journals and Conferences of the Ethics Domain

Appendix D: Manual Query

The following *five* leading journals and conferences in table D.1 are not covered through the databases from our initial keyword-based search. Hence, they are queried manually by searching for our

keywords in the publication titles of the published articles in those nine leading journals or conferences. If an article seems to be relevant but does not include one of the keywords in the publication title, we checked additionally for the keywords in the abstract. Table D.1 provides an overview of the articles identified through the manual keyword search. The identified article undergoes the same screening procedure (format screening, quality screening, and relevance screening) as the articles found in the seven databases.

Research Domain	Outlet	Hits based on Keywords (Publication Titles)	Hits based on Keywords (Abstract)
IS-domain	Conference on Innovation Data Systems Research (CIDR)	0	0
Management and Organization Science-domain	Australian Journal of Labour Law	0	0
	AoM Organizational Communication and Information Systems	0	0
	International Conference in Business Process Management	0	0
Ethics-domain	Journal for Business, Economics & Ethics	0	1

Table D.1: Manual Search Results

Appendix D: Keyword-Based Search Strategy

The keyword field is not available for the databases *ACM Library*, *AIS Electronic Library*, and *IEEE Computer Society Digital Library* (abstract and title only). We did not use wildcards in the following databases: *AIS Electronic Library*, *ScienceDirect*, and *JSTOR*. The keyword-based search strategy (with wildcards) is given in table E.1.

Keyword-Strings	Human-AI Interaction		Strategic Management		Ethical Principles
keywords	substitution OR	AND	“corporate strategy” OR “business strategy” OR “functional strategy” OR “planning process*” OR “strategic process*” OR “tactical process*” OR “operational process*”	AND	transparency OR
	augmentation OR				accountability OR
	assemblage				privacy OR maleficence OR justice OR fairness OR beneficence OR sustainability OR responsibility OR autonomy
	OR		OR		OR

Overarching terms = keywords	“Artificial Intelligence”		manag* OR strateg*		Ethic*

Table E.1: Keyword-Based Search Strategy

We have the following search string (without wildcards) applied to the title, abstract, and keywords (if available):

(“Artificial Intelligence” OR “substitution” OR “augmentation” OR “assemblage”) AND (“ethics” OR “transparency” OR “accountability” OR “privacy” OR “maleficence” OR “justice” OR “fairness” OR “beneficence” OR “sustainability” OR “responsibility” OR “autonomy”) AND (“management” OR “strategy” OR “corporate strategy” OR “business strategy” OR “functional strategy” OR “planning processes” OR “strategic processes” OR “tactical processes” OR “operational processes”).

Appendix F: Format Screening

We exclude articles fulfilling the exclusion criteria given in table F.1. In total, the number of articles that have been deleted through filter options given through some of the seven databases is 3,105 articles. The number of articles that have been deleted through manual format screening is 2,379 articles. Together 5,484 articles have been excluded due to format criteria. Please note that the information given in table F.1 only refers to the keyword-search strategy in the databases and the manual search and not to the backward and forward searches.

Exclusion Criteria	Exclusion Description	Number of Eliminated Articles
Non-scientific articles (n=1,503 articles)	Gray literature, book chapters, book reviews, dissertations, tutorials, bachelor’s or master’s theses, workshops, posters, letter of editors, special issues, symposiums, talks, magazine, and articles of non-scientific conferences.	<u>Through filter options:</u> (total: 0 articles) No filter available <u>Through manual screening:</u> (total: 1,503 articles) ACM: 12 articles, AIS: 1,275 articles, EBSCO: 11 articles, IEEE: 135 articles, JSTOR: 1 article, ScienceDirect: 35 articles, Web of Science: 34 articles
Incomplete articles (n=113 articles)	Short papers or research-in-progress papers, extended abstracts.	<u>Through filter options:</u> (total: 0 articles) No filter available <u>Through manual screening:</u> (total: 113 articles)

		ACM: 3 articles, AIS: 83 articles, EBSCO: 4 articles, IEEE: 4 articles, JSTOR: 1 article, ScienceDirect: 5 articles, Web of Science: 13 articles
Non-peer-reviewed articles (n=669 articles)	Articles not being peer-reviewed.	<u>Through filter options: (total: 417 articles)</u> ACM: 13 articles, AIS: 4 articles, EBSCO: 392 articles, IEEE: 8 articles <u>Through manual screening: (total: 252 articles)</u> ACM: 4 articles, AIS: 28 articles, EBSCO: 7 articles, IEEE: 20 articles, ScienceDirect: 167 articles, Web of Science: 26 articles
Non-English articles (n=502 articles)	Articles not completely written in English are excluded.	<u>Through filter options: (total: 385 articles)</u> EBSCO: 3425 articles, IEEE: 1 article, JSTOR: 37 articles, Web of Science: 2 articles <u>Through manual screening: (total: 117 articles)</u> AIS: 15 articles, EBSCO: 2 articles, JSTOR: 1 article, ScienceDirect: 97 articles, Web of Science: 1 article, manual search: 1 article
Non-abstract articles (n=2,304 articles)	Articles lacking an abstract are excluded due to the mechanism of applying exclusion and inclusion criteria on title, abstract, and keywords.	<u>Through filter options: (total: 2,303 articles)</u> ACM: 63 articles, AIS: 45 articles, EBSCO: 525 articles, JSTOR: 49 articles, Web of Science: 1,621 articles <u>Through manual screening: (total: 1 article)</u> AIS: 1 article
Review articles (n=393 articles)	Standalone literature review papers are excluded.	<u>Through filter options: (total: 0 articles)</u> No filter available <u>Through manual screening: (total: 393 articles)</u> ACM: 1 article, AIS: 222 articles, EBSCO: 30 articles, IEEE: 19 articles, JSTOR: 2 articles, ScienceDirect: 51 articles, Web of Science: 68 articles
<i>Total: 5,484 articles</i>	<i><u>Through filter options: 3, 105 articles</u></i>	<i><u>Through manual screening: 2,379 articles</u></i>

Table F.1: Format Screening

Appendix G: Quality Screening

We only include journal articles published in the Harzing, 2020 journal list (67th ed.; Harzing, 2020)

that fulfill the criteria:

- ranking of C or higher in the ABDC (Australian Business Deans Council Journal Rankings List 2019) *OR*
- ranking of C or higher in the VHB (VHB-JOURQUAL3) *OR*
- ranking of 1 or lower in the Ejis & EjisCI (the European Journal of Information Systems List of 2007 of Mingers and Harzing, 2007) *OR*
- ranking of 4 or higher in the UQ (University of Queensland Adjusted ERA Rankings List 2011)

Furthermore, we include all journal articles that are either part of the top21 list of Lowry et al. (2013) or part of the FT 50 journal list (Financial Times, 2016) (in specific, 13 journal outlets not being published in the Harzing, 2020 journal list (67th ed.; Harzing, 2020): e.g., Organization Science).

We include articles from the proceedings of conferences that were ranked C or higher in the ERA conference ranking list (Australian Research Council, 2011) and/or were ranked D or higher in the VHB-conference ranking list (VHB, 2011). We do not apply quality criteria for ethical outlets.

In total, 367 conference articles and 1,491 journal articles have been removed. The eliminated articles (n=1,858) refer to a total of 1,204 outlets. Thereby, 1,003 journal outlets and 201 conference outlets have been excluded due to the above-mentioned criteria. Table G.1 gives an overview of the outlets being excluded.

Ranking Criteria	Examples of Excluded Outlets	Number of Excluded Outlets
Journal articles not fulfilling the above-mentioned criteria from the Harzing, 2020 journal list (67 th ed.; Harzing, 2020) or are not included in the top21 journal list of Lowry et al. (2013) or in the FT 50 journal list (Financial Times, 2016) being excluded.	Advances in Business-Related Scientific Research Journal, European Journal of Communication, International Journal of Clinical Leadership, Journal of Global Information Technology Management, The Journal of Socio-Economics, etc.	Total: 1,003 articles <u>Distribution:</u> ACM: 8 articles, AIS: 312 articles, EBSCO: 82 articles, IEEE: 6 articles, JSTOR: 1 article, ScienceDirect: 276 articles, Web of Science: 318 articles
Conference articles having a ranking of at least C or higher in the ERA conference ranking list (Australian Research Council, 2011) or have a ranking of at least D or higher in the VHB-conference ranking list (VHB, 2011).	Conference on Emerging Technologies & Factory Automation (ETFA), International Conference on Arts and Culture, International Conference on Multi Agent Systems, International Conference on Technical Debt, Wuhan International Conference on e-Business, etc.	Total: 201 articles <u>Distribution:</u> ACM: 3 articles, AIS: 88 articles, EBSCO: 16 articles, IEEE: 2 articles, ScienceDirect: 85 articles, Web of Science: 7

Table G.1: Quality Screening**Appendix H: Relevance Screening**

To be considered relevant, the articles must meet all three criteria given in table H.1. We measure interrater agreement by using Cohen's Kappa following Cohen (1960) to increase review reliability. The intercoder agreement of parallel relevance screening is $\alpha = 0.81$ (0.05 alpha-level, $n=133$). This indicates a strong agreement between coders (McHugh, 2012). In total, 2,121 articles have been eliminated through relevance screening. Please note that the exclusion criteria for relevance screening may apply to more than one criterion.

Nr.	Relevance Criteria	Number of Excluded Articles
1.	The authors must provide their own conceptualization or understanding of ethics regarding human-AI interaction (provide dedicated contributions to our research subject), and any ethical issues or principles should concern the context of human-AI interaction.	Total: 980 articles <u>Distribution:</u> ACM: 5 articles, AIS: 295 articles, EBSCO: 55 articles, IEEE: 33 articles, JSTOR: 4 articles, ScienceDirect: 531 articles, Web of Science: 57 articles
1 & 2	Combination of criteria 1 and 2	Total: 10 articles <u>Distribution:</u> AIS: 6 articles, EBSCO: 1 article, IEEE: 1 article, ScienceDirect: 2 articles
2.	Technologies mentioned in the articles must carry out tasks that typically require human cognition, following our underlying definition of AI based on Rai et al. (2019). Hence, we included articles that do not claim the term "AI" specifically, in case they provide ideas on independent and self-learning technologies, such as expert systems, being early forms of intelligent technologies with respect to our thoughts on sociomateriality.	Total: 806 articles <u>Distribution:</u> AIS: 464 articles, EBSCO: 56 articles, IEEE: 24 articles, ScienceDirect: 256 articles, Web of Science: 6 articles
2 & 3	Combination of criteria 2 and 3	Total: 5 article <u>Distribution:</u> AIS: 1 article, ScienceDirect: 4 articles
3.	Articles must contain ideas on the strategic management of human-AI interaction in an organizational context having employees as the human part. For instance, the articles must explicitly mention corporate, business, or functional strategies or the associated processes of strategic management of human-AI interaction in organizations or contain ideas that are transferable to the strategic management of human-AI interaction in an organization.	Total: 269 articles <u>Distribution:</u> ACM: 2 articles, AIS: 177 articles, EBSCO: 23 articles, IEEE: 5 articles, ScienceDirect: 36 articles, Web of Science: 26 articles

3 & 1	Combination of criteria 3 and 1	Total: 4 articles <u>Distribution:</u> AIS: 13 articles, EBSCO: 2 articles, ScienceDirect: 17 articles, Web of Science: 10 articles
1, 2 & 3	Combination of all criteria	Total: 8 articles <u>Distribution:</u> ScienceDirect: 6 article, Web of Science: 2 articles

Table H.1: Relevance Screening

Appendix I: Backward and Forward Search

Besides conducting backward and forward searches for the references of the remaining 20 articles of the *final sample* based on the keyword search in the databases and the manual search, we also conducted a backward and forward search for the standalone literature review articles and research-in-progress articles, which we excluded through format criteria during the screening process.

Following the references of the 20 articles of the *final sample* from our keyword-based search strategy (keyword sample), we analyzed 177 additional articles (142 articles in the first round, 32 articles in the second round, and 3 articles in the third round). In addition, following the references of the 393 standalone literature reviews (review sample), we analyzed 447 additional articles (244 articles in the first round, 185 articles in the second round, and 18 articles in the third round). Furthermore, following the references of the 113 research-in-progress articles (RIP sample), we analyzed 276 additional articles (246 articles in the first round and 30 articles in the second round). This resulted in 900 articles being analyzed additionally through backward and forward searches.

In total, we added 22 articles to the *final sample* through backward and forward search. Thereupon 7 articles came from the keyword-based search sample, 8 articles from the standalone literature reviews, and 7 articles from the research-in-progress articles. See table I.1 for an overview:

Search Type	Round	Keyword Sample		Review Sample		RIP Sample	
		Nbr. of Articles	Screening	Nbr. of Articles	Screening	Nbr. of Articles	Screening

Back-ward	1 st round	• 99 articles • Thereof, 9 duplicates	• Excluded format: 45 articles • Excluded quality: 9 articles • Excluded relevance: 31 articles • 5 articles remain	• 179 articles • Thereof, 25 duplicates	• Excluded format: 39 articles • Excluded quality: 70 articles • Excluded relevance: 42 articles • 3 articles remain	• 212 articles • Thereof, 21 duplicates	• Excluded format: 37 articles • Excluded quality: 63 articles • Excluded relevance: 84 articles • 5 articles remain
For-ward	1 st round	• 39 articles • Thereof, 4 duplicates	• Excluded format: 18 articles • Excluded quality: 6 articles • Excluded relevance: 11 articles • 0 articles remain	• 65 articles • Thereof, 14 duplicates	• Excluded format: 20 articles • Excluded quality: 21 articles • Excluded relevance: 8 articles • 2 articles remain	• 34 articles • Thereof, 7 duplicates	• Excluded format: 18 articles • Excluded quality: 3 articles • Excluded relevance: 5 articles • 1 article remain
1 st round total		142 articles being analyzed (5 articles added to the final sample)		244 articles being analyzed (5 articles added to the final sample)		246 articles being analyzed (6 articles added to the final sample)	
Back-ward	2 nd round	• 27 articles • Thereof, 1 duplicate	• Excluded format: 9 articles • Excluded quality: 8 articles • Excluded relevance: 8 articles • 1 article remain	• 77 articles • Thereof, 17 duplicates	• Excluded format: 4 articles • Excluded quality: 29 articles • Excluded relevance: 27 articles • 0 articles remain	• 23 articles • Thereof, 2 duplicates	• Excluded format: 5 articles • Excluded quality: 9 articles • Excluded relevance: 7 articles • 0 articles remain
For-ward	2 nd round	• 5 articles • Thereof, 0 duplicates	• Excluded format: 0 articles • Excluded quality: 3 articles	• 108 articles • Thereof, 14 duplicates	• Excluded format: 26 articles • Excluded quality: 34 articles	• 7 articles • Thereof, 1 duplicate	• Excluded format: 1 article

			- Excluded relevance: 1 article 1 article remain		- Excluded relevance: 31 articles 3 articles remain		- Excluded quality: 4 articles - Excluded relevance: 0 articles 1 article remain
2 nd round total		32 articles being analyzed (2 articles added to the final sample)		185 articles being analyzed (3 articles added to the final sample)		30 articles being analyzed (1 article added to the final sample)	
Backward	3 rd round	3 articles - Thereof, 0 duplicates	- Excluded format: 0 articles - Excluded quality: 0 articles - Excluded relevance: 3 articles 0 articles remain	8 articles - Thereof, 2 duplicates	- Excluded format: 2 articles - Excluded quality: 1 article - Excluded relevance: 3 articles 0 articles remain	4 articles - Thereof, 1 duplicate	- Excluded format: 0 articles - Excluded quality: 2 articles - Excluded relevance: 1 article 0 articles remain
Forward	3 rd round	-	-	10 articles - Thereof, 5 duplicates	- Excluded format: 1 article - Excluded quality: 1 article - Excluded relevance: 3 articles 0 articles remain	3 articles - Thereof, 1 duplicate	- Excluded format: 1 article - Excluded quality: 1 article - Excluded relevance: 0 articles 0 articles remain
3 rd round total		3 articles being analyzed (0 articles added to the final sample)		18 articles being analyzed (0 articles added to the final sample)		7 articles being analyzed (0 articles added to the final sample)	
Total: 863 analyzed articles		171 analyzed articles (7 articles added to the final sample)		423 analyzed articles (8 articles added to the final sample)		276 analyzed articles (7 articles added to the final sample)	

Table I.1: Backward and Forward Search

Appendix J: Final Sample

The *final sample* (n=42) consists of 18 conference articles and 24 journal articles. Thereby, all conference articles have their origin in the IS domain. Most of the journal articles are also from the IS domain (n=15), whereby there are also 2 journal articles from the Ethics domain and 7 journal articles from the Management and Organization Science domain in the *final sample*. While 20 articles of the *final sample* are from our keyword-based search strategy in the seven databases and manual search, 22 articles are from backward and forward searches. An overview of the *final sample* is given in table J.1:

Nbr.	Autor(s)	Year	Title	Outlet	Source
1	Acikgoz et al.	2020	Justice perceptions of artificial intelligence in selection	International Journal of Selection & Assessment	Keyword sample
2	Asatiani et al.	2020	Challenges of explaining the behavior of black-box AI systems	MIS Quarterly Executive	Keyword sample
3	Bankins et al.	2022	AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context	Information Systems Frontiers	Review sample (forward 2 nd round)
4	Bansal et al.	2021	Does the whole exceed its parts? The effect of AI explanations on complementary team performance	Conference on Human Factors in Computing Systems	RIP sample (backward 1 st round)
5	Basu et al.	2021	Human decision making in AI augmented systems: evidence from the initial coin offering market	Hawaii International Conference on System Sciences (HICCS)	RIP sample (backward 1 st round)
6	Cranefield & Doyle	2022	“You have a few wordy spots”: role-related impacts of working with an intelligent assistant	European Conference on Information Systems (ECIS)	Keyword sample
7	Coombs et al.	2021	What is it about humanity that we can't give away to intelligent machines? A European perspective	International Journal of Information Management	Review sample (forward 2 nd round)
8	Ehsan & Riedl	2022	Requirements for AI-based teammates: a qualitative inquiry in the context of creative workshops	Hawaii International Conference on System Sciences (HICCS)	RIP sample (forward 1 st round)
9	Figueroa-Armijos et al.	2022	Ethical perceptions of AI in hiring and organizational trust: the role of performance expectancy and social influence	Journal of Business Ethics	Review sample (forward 2 nd round)

10	Fuegener et al.	2021	Cognitive challenges in human-AI collaboration: investigating the path towards productive delegation	Information Systems Research	RIP sample (backward 1 st round)
11	Grønsund & Aanestad	2022	Augmenting the algorithm: emerging human-in-the-loop work configurations	The Journal of Strategic Information Systems	Review sample (backward 1 st round)
12	Gupta et al.	2021	Role of risks in the development of responsible Artificial Intelligence in the Digital Healthcare Domain	Information Systems Frontiers	Keyword sample
13	Hinsen et al.	2022	How can organizations design purposeful human-AI interactions: a practical perspective from existing use cases and interviews	Hawaii International Conference on System Sciences (HICCS)	Review sample (forward 1 st round)
14	Hoeddinghaus et al.	20	The automation of leadership functions: Would people trust decision algorithms?	Computers in Human Behavior	Review sample (backward 1 st round)
15	Hohenstein & Jung	2020	AI as a moral crumple zone: the effects of AI-mediated communication on attribution and trust	Computers in Human Behavior	Keyword sample
16	Holstein et al.	2019	Improving fairness in machine learning systems: what do industry practitioners need?	Conference on Human Factors in Computing Systems	RIP sample (backward 1 st round)
17	Jia et al.	2022	Do we blame it on the machine? Task outcome and agency attribution in human-technology collaboration	Hawaii International Conference on System Sciences (HICCS)	Review sample (forward 1 st round)
18	Jiang et al.	2021	Algoactivistic practices in ridesharing - a topic modeling & grounded theory approach	European Conference on Information Systems (ECIS)	Keyword sample
19	Kelley	2022	Employee perceptions of the effective adoption of AI principles	Journal of Business Ethics	Keyword sample
20	Klumpp	2018	Automation and artificial intelligence in business logistics systems: human reactions and collaboration requirements	International Journal of Logistics Research and Applications	Keyword sample (backward 1 st round)

21	Klumpp & Zijm	2019	Logistics innovation and social sustainability: how to prevent an Artificial Divide in human-computer interaction	International Journal of Logistics Research and Applications	Keyword sample
22	Kolbjørnsrud et al.	2017	Partnering with AI: how organizations can win over skeptical managers	International Journal of Logistics Research and Applications	Keyword sample (backward 2 nd round)
23	Lee	2018	Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management	Big Data & Society	Keyword sample (forward 2 nd round)
24	Lee et al.	2015	Working with machines: the impact of algorithmic and data-driven management on human workers	International Conference on Computer-Human Interaction	Keyword sample (backward 1 st round)
25	Liao et al.	2020	Questioning the AI: informing design practices for explainable AI user experiences	Conference on Human Factors in Computing Systems	RIP sample (backward 2 nd round)
26	Madaio et al.	2020	Co-designing checklists to understand organizational challenges and opportunities around fairness in AI	International Conference on Computer-Human Interaction	Keyword sample (backward 1 st round)
27	Malik et al.	2022	Impact of artificial intelligence on employees working in industry 4.0 led organizations	International Journal of Logistics Research and Applications	Keyword sample
28	Mayer et al.	2021	How corporations encourage the implementation of AI ethics	European Conference on Information Systems (ECIS)	Keyword sample
29	Mayer et al.	2020	Unintended consequences of introducing AI systems for decision making	MIS Quarterly Executive	Keyword sample
30	Mirbabaie et al.	2020	The rise of Artificial Intelligence – understanding the AI identity threat at the workplace	Electronic Markets	Keyword sample (backward 1 st round)
31	Park et al.	2021	Human-AI interaction in Human Resource Management: understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens	Conference on Human Factors in Computing Systems	Keyword sample

32	Pregenzer et al.	2021	Algorithms in the driver's seat: explaining workers' reactions to algorithmic control	European Conference on Information Systems (ECIS)	Keyword sample
33	Polyviou & Zamani,	2022	Are we Nearly There Yet? A Desires & Realities Framework for Europe's AI Strategy	Information System Frontiers	Keyword sample
34	Rajaobelina et al.	2021	Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot	Psychology & Marketing	Review sample (backward 1 st round)
35	Rix & Hess	2022	Hello, mate! Insights from the field on leveraging machine teammates in organizations	Pacific Asia Conference on Information Systems (PACIS)	Keyword sample
36	Sauppé & Mutlu	2015	The social impact of a robot co-worker in industrial settings	Conference on Human Factors in Computing Systems	RIP sample (backward 1 st round)
37	Shin & Park	2019	Role of fairness, accountability, and transparency in algorithmic affordance	Computers in Human Behavior	Keyword sample (backward 1 st round)
38	Someh et al.	2022	Building an Artificial Intelligence explanation capability	MIS Quarterly Executive	Keyword sample
39	Strich et al.	2021	What do I do in a world of Artificial Intelligence? Investigating the impact of substitutive decision-making AI systems on employee's professional role identity	Journal of the Association for Information Systems	Keyword sample
40	Wang et al.	2020	Toward an understanding of responsible Artificial Intelligence practices	Hawaii International Conference on System Sciences (HICSS)	Keyword sample
41	Watson et al.	2021	Will AI ever sit at the C-suite table? The future of senior leadership	Business Horizons	Keyword sample
42	Zhang et al.	2020	Developing fairness rules for talent intelligence management system	Hawaii International Conference on System Sciences (HICSS)	Keyword sample

Table J.1: Final Sample

Appendix K: Meta Information

Table K.1 provides an overview of the *meta information* of the *final sample*:

Meta Information	Codes	Number of Articles	Article Nbr.
Research Discipline	IS Domain	33 articles	2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 23, 24, 25, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 41, 42
	Management and Organization Science Domain	7 articles	1, 20, 21, 22, 27, 34, 40
	Ethics Domain	2 articles	9, 19
Database	ACM	8 articles (thereof, 1 article is also in the Web of Science database)	4, 16, 24, 25, 26, 31, 36
	AIS	13 articles (thereof, 2 article are also in other databases)	2, 6, 13, 17, 18, 28, 29, 32, 35, 38, 39, 40, 42
	EBSCO	3 articles	19, 21, 27
	IEEE	1 article (this article is also in the AIS and ScienceDirect database)	32
	JSTOR	0 articles	-
	ScienceDirect	8 articles (thereof, 2 articles are also in other databases)	1, 5, 6, 8, 11, 12, 15, 32, 41
	Web of Science	22 articles (thereof, 1 article is also in the ACM database)	31, 33
	Other databases (articles from the backward and forward search)	11 articles (Elsevier, Sagehub, Springer, SSRN, Wiley Online Library)	3, 7, 9, 10, 14, 20, 22, 23, 30, 34, 37
Publication Type	Journal	24 articles	1, 2, 3, 7, 9, 10, 11, 12, 14, 15, 19, 20, 21, 22, 23, 26, 27, 29, 33, 34, 37, 38, 39, 41
	Conference	18 articles	4, 5, 6, 8, 13, 16, 17, 18, 24, 25, 26, 28, 31, 32, 35, 36, 40, 42
Outlet Ranking	High (e.g., A*, A)	2 articles	11, 40

	Middle (e.g., B, C)	32 articles	2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 30, 32, 33, 34, 35, 37, 38, 40, 41, 42
	Low (e.g., D)	8 articles	1, 4, 16, 24, 25, 26, 31, 36
Paper Type	Empirical	42 articles	1-42
	Non-empirical	0 articles	-
Method Type I	Qualitative	26 articles	2, 6, 7, 8, 11, 13, 18, 19, 20, 21, 23, 24, 26, 27, 28, 29, 31, 32, 33, 35, 36, 38, 39, 40, 41, 42
	Quantitative	10 articles	1, 5, 9, 10, 12, 14, 15, 17, 34, 37
	Mixed-Method	6 articles	3, 4, 16, 22, 25, 30
Method Type II	Online Survey	9 articles	3, 9, 12, 15, 16, 22, 23, 30, 34
	Interviews	14 articles	6, 7, 13, 18, 19, 24, 25, 26, 27, 31, 35, 36, 37, 41, 42
	Case Study	11 articles	2, 5, 8, 11, 20, 21, 28, 29, 38, 39, 40
	Experiment	5 articles	1, 4, 10, 14, 17
	Delphi Study	1 article	42
	Social Media / Text Analysis	2 articles	32, 33
Analysis Type	Qualitative Analysis	20 articles	6, 7, 13, 19, 20, 21, 24, 25, 26, 27, 28, 29, 31, 33, 35, 38, 39, 40, 41, 42
	Descriptive Statistics	9 articles	1, 9, 10, 12, 14, 15, 17, 34, 37
	Hybrid Analysis Forms	13 articles	2, 3, 4, 5, 8, 11, 16, 18, 22, 23, 30, 32, 36
Sample Type	Students	2 articles	1 (also an employee sample), 17
	Employees/Workers	18 articles	1 (also a student sample), 2, 4, 9, 10, 12, 18, 19, 20, 21, 23, 24, 26, 27, 28, 31, 32, 34
	Managers	6 articles	11, 22, 36, 38, 39, 41
	Researchers	1 article	7
	Experts	8 articles	5, 6, 13, 25, 28, 29, 34, 42
	No Sample (or not claimed)	2 articles	16, 40
	Random Sample (citizens, users)	5 articles	3, 8, 14, 15, 33, 37
Sample Size	<100	22 articles	2, 6, 7, 8, 13, 17, 19, 20, 21, 24, 25, 26, 27, 28, 29, 31, 35, 36, 39, 41, 42
	100-299	5 articles	9, 12, 15, 18, 23
	299-599	7 articles	1, 14, 16, 30, 33, 34, 37
	600-1000	3 articles	3, 10, 32
	>1000	3 articles	4, 5, 22

	No sample (or not claimed)	2 articles	16, 40
Result Type	Propositions / Future Research Directions	2 articles	15, 16
	Framework / Classification / Summary	10 articles	2, 7, 8, 11, 18, 23, 27, 33, 39, 42
	Hypotheses-Validation	9 articles	1, 3, 4, 5, 9, 10, 30, 32
	Theory-Development	7 Articles	6, 14, 17, 20, 21, 32, 35
	(Practical/Design) Guidelines	13 articles	13, 19, 22, 24, 25, 26, 28, 29, 31, 36, 38, 40, 41
Contribution Type	(rather) theoretical	14 articles	5, 6, 7, 8, 10, 14, 15, 16, 18, 27, 30, 32, 34, 35
	(rather) practical	12 articles	2, 13, 19, 22, 28, 29, 31, 33, 36, 38, 40, 41
	Both	15 articles	1, 3, 9, 11, 12, 17, 20, 21, 23, 24, 25, 26, 37, 39, 42
	Not claimed	1 article	4
Use of Theory	Exploratory	21 articles	2, 8, 11, 13, 15, 16, 17, 23, 24, 25, 26, 31, 32, 35, 36, 37, 38, 39, 40, 41, 42
	Deductive	12 articles	1, 5, 9, 10, 12, 14, 19, 20, 21, 27, 30, 31
	Inductive	7 articles	6, 7, 18, 22, 28, 29, 33
	Mixed	2 articles	3, 4

Table K.1: Meta Information of the Final Sample

Appendix L: Coding Schema

For deductive coding, we follow the steps of Mayring (2014, p. 95): 1. Definition of the research question and the theoretical background 2. Definition of the category system (main categories and subcategories, 3. Definition of coding guidelines, 4. A material run-through, preliminary coding, coding rules, 5. Revision of the categories and coding guidelines after 10-50 % of the material, 6. Final working through the material, and 7. Analysis, category frequencies, and contingencies interpretation. Whereas the first step is already done before (see sections 1, 2, and 4), the remaining steps are part of section 5. The deductive coding scheme is mainly based on section 2 (theoretical lens of sociomateriality and sociotechnical system theory) and section 4 (theoretical background on the main constructs). Table J.1 provides a snapshot of the deductive coding scheme:

Category	Definition	Sub-category	Definition	Coding Rule	Example
Inseparability	Inseparability describes the multiple ways social and material entities can be interwoven (Jones, 2014). It focuses on the interaction process between the social and the material entity and the role each entity plays for the other during the interaction (limited to the imbrication during the process of fulfilling a specific task or goal). The notion suggests that neither entity can function separately. Although Orlikowski and Scott (2008, p. 434) suggest an “inherent inseparability” of the social and material entity, inseparability is not restricted to a specific ontology, as the interrelationship of both entities can vary between strong and weak (Jones, 2014).	Strong inseparability	Strong inseparability describes a “mutual constitution” (Jones, 2014, p. 920). The material and social entities are one unit or a symbiosis, given that certain tasks or functions do not work without the other entity.	The text must contain at least one of the criteria: - Inter dependencies between both entities (certain tasks or functions do not work without the other entity) - Both entities shape each other or form a unit	“Drivers need to turn on their ridesharing app to be able to receive and execute their work” (Lee et al., 2015, p. 1604) OR “Alder Play [an AI-based APP used in hospitals] enables healthcare professionals to have access to medical records of patients who are eligible for NHS treatment. Patients and their families would be able to obtain their medical records online. This could largely improve transparency in the clinical processes, thereby enhancing the quality of health services and strengthening the patient engagement.” (Wang et al., 2020, p. 4965)

Table L.1: Snapshot of the Deductive Coding Scheme

Appendix M: Types of Human-AI Interaction

The three types of human-AI interaction following Rai et al. (2019) are given in table M.1:

Type	Characteristics	Similar terms
Substitution	- AI substitutes human agents in performing a task (Rai et al., 2019) - Replacing the tasks people used to do (Norman, 2017) - Keep humans out of the equation to allow more comprehensive, rational, and efficient processing (Davenport and Kirby, 2016, p. 26)	- Task elimination (Norman, 2017) - Task automation (Raisch and Krakowski, 2020) - Human automation (Janssen et al., 2019) - Human-out-of-the-loop (Shahriari et al., 2016)

	● “automation involves the substitution of machines for labor and leads to the displacement of workers from the tasks that are being automated” (Acemoglu, D., Restrepo, P., 2018, p. 5)	
Augmentation	● The main focus is to balance each other’s weaknesses and learn from each other by using the complementary strengths of human intelligence and AI (Daily et al., 2017; Kamar, 2016; Raisamo et al., 2019) ● <i>Machine augmentation</i>: bringing human intelligence in the AI to design, complement, and evaluate the capabilities of AI (Mnih et al., 2015) or by having the human as a teacher bringing the system to learn (Dellermann et al., 2019b) ● <i>Human augmentation</i>: AI helping humans expand their abilities by extending physical capabilities or interacting with individuals more effectively (Wilson and Daugherty, 2018, p. 214)	● Human-in-the-loop (Daily et al., 2017; Raisamo et al., 2019; Wilson and Daugherty, 2018) ● <i>Human augmentation</i> (Raisamo et al., 2019) ● Augmented Human Intelligence (Dellermann et al., 2019b) ● Augmented Machine Intelligence (Dellermann et al., 2019b)
Hybrid Intelligence	● “systems that have the ability to accomplish complex goals by combining human and artificial intelligence to collectively achieve superior results” (Dellermann et al., 2019a, p. 1) ● “AI and human agents are assembled contextually and temporally ... in assemblage the human and AI agents work as an integrated unit” (Rai et al., 2019, p. iv) ● Distribution of roles in hybrid intelligence: augmented human intelligence and augmented machine intelligence (Dellermann et al., 2019b, p. 640) ● “The performance of Hybrid Intelligence systems can, thus, not only be measured by the superior outcome of the whole sociotechnical system alone, but the learning ... of human and machine agents” (Dellermann et al., 2019b, pp. 640–641) ● “It requires deeper reasoning capabilities for machines to make decisions not only about how they are accomplishing their tasks, but also about how they can support their teammates towards the success of the collaborative activity” (Kamar, 2016, p. 4073)	● Assemblage (Rai et al., 2019) ● Combining augmented human and augmented machine intelligence (Dellermann et al., 2019b)

Table M.1: Types of Human-AI Interaction

Appendix N: Ethical Principles

In 2004, the politically driven attempt by the Organization for Economic Co-operation and Development (OECD), launched in 2004, to concretize ethics for businesses centered on four widely

applicably principles: fairness, transparency, accountability, and responsibility (Rezaee, 2008, p. 35). It should be noted that we chose our own classification based on the most common principles of business ethics we found in business ethics literature, which are given in table N.1:

Ethical principles	Characteristics	Authors
Transparency & understandability (distinctiveness, visibility, accuracy)	• Understandable <i>management processes</i> (Rezaee, 2008) • Enhancing accountability through assurance (Dando and Swift, 2003)	e.g., Dando and Swift, 2003; Rezaee, 2008
Security & (non-)maleficence (human/civil/employment rights, no-harassment, security, safety, political contributions, compliance, non-discrimination, no-risk)	• Avoid conflicts of interest (Payne and Wayland, 1999) • Manage personal biases (Payne and Wayland, 1999) • Comply with law and human rights (Goodpaster, 1991)	e.g., Dando and Swift, 2003; Goodpaster, 1991; Payne and Wayland, 1999; Winstanley et al., 1996
Fairness & justice (consistency, consensus, diversity, dignity, respect, inclusiveness, non-discrimination)	• Treat all people with equal respect (Hasnas, 2013) • Promote fairness and justice for all stakeholders (Rezaee, 2008)	e.g., Hasnas, 2013; Payne and Wayland, 1999; Rezaee, 2008; Winstanley et al., 1996
Sustainability & credibility (relevance, well-being, innovation)	• Enforce long-term impact of management practices (Goodpaster, 1991) • Enhance human well-being (Hasnas, 2013) • Strengthen social, ethical, and environmental credibility (Dando and Swift, 2003)	e.g., Goodpaster, 1991; Hasnas, 2013; Winstanley et al., 1996
Responsibility & leadership (authority, legitimacy, agreement, validity, obligations, credibility, control, deregulation)	• Balance between decision-making power of all organizational stakeholders (Hasnas, 2013; Payne and Wayland, 1999) • Ensure that people act with awareness (Hasnas, 2013) • Increase organizational profits so long as there are no negative consequences for organizational stakeholders (Greenwood, 2002)	e.g., Hasnas, 2013; Greenwood, 2002; Payne and Wayland, 1999; Schwartz and Saiia, 2012; Winstanley et al., 1996

Table N.1: Principles of Business Ethics

Several authors aggregated and categorized ethical principles in the context of AI. For instance, Leslie (2019) identified key attributes of AI ethics, such as fairness, accountability, and transparency. Bostrom and Yudkowsky (2014) claimed criteria of machine ethics related to the AI algorithm, such as responsibility or transparency. Arrieta et al. (2019) identified understandability, comprehensibility, and interpretability, for instance, whereas Fjeld et al. (2020) proposed eight core AI principles, including privacy, security, and so on. Whittlestone et al. (2019) also explored the use of ethical principles in research, with privacy as the word appearing most frequently. In their scoping review, Jobin et al. (2019) identified 11 ethical principles related to guidelines on ethical AI, such as (non-)maleficence. In addition, there are several initiatives concerned with the ethical implications of AI, such as AI4People, the Asilomar AI principles of the Future of Life Institute, the General Principles of IEEE (Institute of Electrical and Electronics Engineers) or the Ethical Principles published by the European Commission's European Group on Ethics in Science and New Technology (Floridi et al., 2018). Synthesizing those principles together in one cluster, we present the principles of AI ethics in table N.2. It should be noted that we chose our own classification based on the most common AI principles named by the authors.

AI principles	Characteristics
Transparency & accountability (explainability, explicability, verifiability, understandability, interpretability, intelligibility, replicability, accuracy)	• Distribute accountability (Fjeld et al., 2020) • Provide information to understand outputs and scope of usage (Fjeld et al., 2020; Floridi et al., 2018) • Avoid non-transparent and unreliable outcomes (Leslie, 2019) • Demonstrate who accounts for the decision-making process and make decisions reviewable (Fjeld et al., 2020; Jobin et al., 2019) • Increase explainability and interpretability (Jobin et al., 2019)
Privacy and (non-) maleficence (security, safety, protection, no harm, precaution, prevention, integrity, incorruptibility, caution, reliability,)	• Avoid misuse, harm, and negative consequences (Floridi et al., 2018; Jobin et al., 2019) • Ensure that individuals' rights are respected (Fjeld et al., 2020; Leslie, 2019) • Avoid unwanted bias (Jobin et al., 2019; Leslie, 2019) • Ensure that AI performs as intended, being safe and resistant against unauthorized users (Fjeld et al., 2020)
Justice and fairness (consistency, inclusion, solidarity, equality, equity, diversity, plurality, accessibility, reversibility, remedy, redress, distribution, prosperity, non-bias, non-discrimination)	• Respect all interests by eliminating discrimination (Floridi et al., 2018) • Promote prosperity and solidarity (Floridi et al., 2018) • Maximize and promote inclusion/inclusiveness (Fjeld et al., 2020) • Provide fair access to AI (Jobin et al., 2019)

Beneficence & sustainability (well-being, dignity, peace, social or common good, benefits, long-term orientation)	● Promote well-being (Floridi et al., 2018) ● Ensure positive benefits (Jobin et al., 2019) and long-term sustainable impact for mankind (Floridi et al., 2018) ● Promote human values (Fjeld et al., 2020) ● Ensure sustainable future developments (Jobin et al., 2019)
Responsibility & autonomy (freedom, consent, choice, self-determination, liberty, empowerment, liability, integrity)	● Ensure that important decisions remain by humans (Fjeld et al., 2020; Floridi et al., 2018) so that humans as the actor who is ultimately responsible (Jobin et al., 2019) ● Ensure that individuals involved in the development and deployment of AI-based technologies are able to evaluate its impact and can be judged based on their professionalism (Fjeld et al., 2020) ● Increase the balance between decision-making power of the <i>human agency</i> and the <i>material agency</i> (Floridi et al., 2018) ● Ensure the protection of the intrinsic value of human choice and human self-determination (Floridi et al., 2018; Jobin et al., 2019; Leslie, 2019)

Table N.2: Principles of AI Ethics

As our focus lies on the ethical managing human-AI interaction from both perspectives: AI interacting with humans (SQ1), but also humans interacting with AI (SQ2), we combine the principles of business ethics and AI ethics into categories of “human-AI ethics” in table N.3:

Categories Human-AI Ethics	Characteristics
Transparency & accountability (accuracy, explainability, explicability, distinctiveness, instrumentality, interpretability, intelligibility, reliability, replicability, understandability, visibility, verifiability)	<i>Duty ethics</i> : principles that ensure that processes of human-AI interaction are easy to understand, decisions are reviewable, and processes of human-AI interaction can be investigated dependent on the interests.
	<i>Virtue ethics</i> : the human’s and AI’s/developer’s ideology that ensures that processes of human-AI interaction are easy to understand, decisions are reviewable, and processes of human-AI interaction can be investigated dependent on the interests.
Privacy and (non-) maleficence (caution, incorruptibility, integrity, no harm, precaution, prevention, protection, safety, security)	<i>Duty ethics</i> : principles that prevent personal privacy and security of humans during human-AI interaction and avoid misuse of AI through humans and vice versa.
	<i>Virtue ethics</i> : the human’s and AI’s/developer’s ideology that aims to prevent personal privacy and security of humans during human-AI interaction and avoid misuse of AI through humans and vice versa.

Justice and fairness (accessibility, consensus, consistency, distribution, diversity, equality, equity, inclusion, non-bias, non-discrimination, plurality, prosperity, redress, relevance, remedy, reversibility, solidarity)	<i>Duty ethics:</i> principles that are based on the respect of all interests of humans to maximize inclusion and the equal treatment of humans in all types of human-AI interaction through adequate management processes.
	<i>Virtue ethics:</i> the human's and AI's/developer's ideology that aims to respect all interests of humans to maximize inclusion and the equal treatment of humans in all types of human-AI interaction through adequate management processes.
Beneficence & sustainability (benefits, dignity, long-term orientation, peace, social or common good, well-being)	<i>Duty ethics:</i> principles that aim for beneficial human-AI interactions with a long-term positive impact.
	<i>Virtue ethics:</i> the human's and AI's/developer's ideology that aims to achieve beneficial human-AI interactions with a long-term positive impact.
Responsibility & autonomy (agreement, authority, consent, choice, empowerment, freedom, legitimacy, liberty, liability, self-determination, validity)	<i>Duty ethics:</i> principles that balance the power of AI and humans in human-AI interaction or managing human-AI interaction as they are based on the agreement among all entities being involved in the human-AI interaction and its consequences.
	<i>Virtue ethics:</i> the human's and AI's/developer's ideology that aims to balance the power of AI and humans in human-AI interaction or to manage human-AI interaction as they are based on the agreement among all entities being involved in the human-AI interaction and its consequences.
Humanity (can be associated with the “uncanny valley” phenomenon in robotics following Park et al. (2021, p. 7) and Rajaobelina et al. (2021, p. 2351))	<i>Duty ethics:</i> principles that aim to consider possible disagreements between humans and AI linked to the transfer of ethical knowledge and experiences between humans and AI by integrating feedback loops or human interfaces for emotional support in the process of human-AI interaction.
	<i>Virtue ethics:</i> the human's and the manager's ideology that aims to use their experiences and emotional nature to integrate humanity, emotional support, and human values in human-AI interaction to reduce moral burden.

Table N.3: Categories of Human-AI Ethics

Appendix O: Strategic Management

As mentioned above, strategic management contains the management of several strategies on three levels: the corporate, the business, and the functional level. The strategies behind those levels serve as a base for the keyword search (see table O.1).

Levels	Description	Authors

Corporate Level	• Overarching corporate or global strategy • Strategic planning and coordination of several substrategies to gain competitive advantages (a rather external view) • Organizations goals, vision, and mission	(Henderson and Venkatraman, 1999; Karimi and Konsynski, 2002, pp. 92–93; Ritson, 2008, p. 13; Somogyi and Galliers, 2009, pp. 11–12)
Business Level	• Development of business strategies • Linked to the organizations' structure and internal processes (a rather internal view)	
Functional Level	• Development of strategies that are related to single departments of the organization	

Table O.1: Levels of Strategic Management

We further identified four processes of strategic management given in table O.2:

Processes	Description	Authors
Planning Processes	• Based on long-term planning horizons • Especially linked to the corporate strategy as it is about making formal strategic plans regarding the corporate strategy on the corporate level	Galliers and Sutherland, 2009, p. 39; Henderson and Venkatraman, 1999; Karimi and Konsynski, 2002, pp. 92–93; Peppard and Ward, 2004; Somogyi and Galliers, 2009, 8, 15
Strategic Processes	• Based on long-term planning horizons • Rather planning activities • Actions mainly on the business level to achieve the several business strategies, but also the corporate strategy (sometimes also on the corporate level)	
Tactic Processes	• Based on middle planning horizons • Coordination and control activities • Actions on the business level to evaluate business strategies and foster functional strategies by determining how best practices can be embedded in operational processes • Between the functional level and the business level	
Operational Processes	• Based on short planning horizons • Actions on the functional level to optimize operations by meeting the budget targets and being productive operatively • Daily working routines to achieve functional strategies	

Table O.2: Processes of Strategic Management

References

- Abdelaal, A., 2021. Grand Research Challenges Facing Ethically Aligned Artificial Intelligence, in: Proceedings of the 27th American Conference on Information Systems (AMCIS). Proceedings of the 27th American Conference on Information Systems (AMCIS 2021), Virtual. 09.08.2021-13.08.2021.
- Acemoglu, D., Restrepo, P., 2018. Artificial Intelligence, Automation and Work 24196, 43 pp. <http://www.nber.org/papers/w24196> (accessed 27 January 2022).
- Acikgoz, Y., Davison, K.H., Compagnone, M., Laske, M., 2020. Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment (IJSA)* 28, 399–416. <https://doi.org/10.1111/ijsa.12306>.
- Alsheibani, S.A., Messom, D., Cheung, Y., Alhosni, M., 2020. Reimagining the Strategic Management of Artificial Intelligence: Five recommendations for business leaders, in: Proceedings of the 26th American Conference on Information Systems (AMCIS 2020). 26th Americas Conference on Information Systems (AMCIS 2020), Salt Lake City, Utah, USA. 10.08.2020-14.08.2020, pp. 1–10.
- Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., 2019. Guidelines for Human-AI Interaction, in: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI). CHI '19: CHI Conference on Human Factors in Computing Systems (CHI), Glasgow, Scotland. 04.05.2019.-09.05.2019, pp. 1–13.
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2019. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, 67 pp. <https://arxiv.org/pdf/1910.10045v2.pdf> (accessed 27 January 2020).
- Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A., 2020. Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive (MISQE)* 19, 259–278. <https://doi.org/10.17705/2msqe.00037>.
- Association for Information Systems, 2011. Senior scholars' basket of journals. <https://aisnet.org/page/SeniorScholarBasket> (accessed 16 January 2022).
- Australian Research Council, 2011. Excellence in research for Australia 2010 ranked conference list. https://webarchive.nla.gov.au/awa/20110217201708/http://www.arc.gov.au/pdf/ERA_report.pdf (accessed 27 January 2022).
- Baidoo-Anu, D., Owusu Ansah, L., 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Available at SSRN 4337484.
- Baird, A., Maruping, L.M., 2021. The next generation of research on IS use: A theoretical framework of delegation to and from agentics IS artifacts. *MIS Quarterly (MISQ)* 45, 315–341. <https://doi.org/10.25300/MISQ/2021/15882>.

- Bankins, S., Formosa, P., Griep, Y., Richards, D., 2022. AI decision making with dignity? Contrasting workers' justice perceptions of Human and AI decision making in a Human Resource Management context. *Information Systems Frontiers (ISF)* -, 1–19. <https://doi.org/10.1007/s10796-021-10223-8>.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance, in: *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21: CHI Conference on Human Factors in Computing Systems (CHI 2021), Yokohama, Japan. 08.05.2021-13.05.2021. Association for Computing Machinery, New York, NY, USA, pp. 1–16.
- Barad, K., 2003. Posthumanist Performativity: Toward an understanding of how matter comes to matter. *Signs: Journal of Women in Culture and Society* 28, 801–831. <https://doi.org/10.1086/345321>.
- Baron, L.F., Gomez, R., 2016. The Associations between Technologies and Societies: The Utility of Actor-Network Theory. *Science, Technology and Society* 21, 129–148. <https://doi.org/10.1177/0971721816640615>.
- Basu, S., Garimella, A., Han, W., Dennis, A., 2021. Human decision making in AI augmented systems: Evidence from the initial coin offering market, in: *Proceedings of the 54th Hawaii International Conference on Systems (HICCS)*. 54th Hawaii International Conference on System Sciences (HICSS-54), Kauai, Hawaii. 05.01.2021, pp. 176–185.
- Benbya, H., Davenport, T.H., Pachidi, S., 2020. Editors' comment: Artificial Intelligence in organizations: current state and future opportunities. *MIS Quarterly Executive (MISQE)* 19, 9–21. <https://doi.org/10.2139/ssrn.3741983>.
- Berente, N., Gu, B., Recker, J., Santhanam, R., 2021. Managing Artificial Intelligence. *MIS Quarterly (MISQ)* 45, 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>.
- Boada, J.P., Maestre, B.R., Genís, C.T., 2021. The ethical issues of social assistive robotics: A critical literature review. *Technology in Society* 67, 101726, 1–13. <https://doi.org/10.1016/j.techsoc.2021.101726>.
- Boell, S.K., Cecez-Kecmanovic, D., 2015. On being 'systematic' in literature reviews in IS. *Journal of Information Technology (JIT)* 30, 161–173. <https://doi.org/10.1057/jit.2014.26>.
- Bonn, I., Fisher, J., 2005. Corporate Governance and Business Ethics: Insights from the strategic planning experience. *Corporate Governance: An International Review* 13, 730–738. <https://doi.org/10.1111/j.1467-8683.2005.00466.x>.
- Bostrom, N., Yudkowsky, E., 2014. The ethics of Artificial Intelligence, in: Frankish, K., Ramsey, W.R. (Eds.), *The Cambridge handbook of Artificial Intelligence*. Cambridge University Press, UK, Cambridge, pp. 316–334.
- Bowie, N.E., 1998. A kantian theory of meaningful work. *Journal of Business Ethics (JBE)* 17, 1083–1092. <https://doi.org/10.1023/A:1006023500585>.
- Burton-Jones, A., Butler, B.S., Scott, S., Xu, S.X., 2021. Next-generation information systems theorizing: A call to action. *MIS Quarterly (MISQ)* 45, 301–314. <https://doi.org/10.25300/MISQ/2021/15434>.
- Carter, L., Liu, D., Cantrell, C., 2020. Exploring the intersection of the digital divide and Artificial Intelligence: A hermeneutic Literature Review. *AIS Transactions on Human-Computer Interaction (THCI)* 12, 253–275. <https://doi.org/10.17705/1thci.00138>.

- Cecez-Kecmanovic, D., Galliers, R.D., Henfridsson, O., Newell, S., Vidgen, R., 2014. The Sociomateriality of Information Systems: Current Status, Future Directions. *MIS Quarterly (MISQ)* 38, 809–830. <https://doi.org/10.25300/MISQ/2014/38:3.3>.
- Chao, C.-H., 2019. Ethics issues in Artificial Intelligence, in: 24th International Conference on Technologies and Applications of Artificial Intelligence (TAAI 2019). 24th International Conference on Technologies and Applications of Artificial Intelligence (TAAI 2019), Kaohsiung, Taiwan. 21.11.2019-23.11.2019, pp. 1–6.
- Churchill, L.R., 1999. Are we professionals? A critical look at the social role of bioethicists. *Daedalus* 128, 253–274.
- Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 37–46.
- Coldwell, D.A., Billsberry, J., van Meurs, N., Marsh, P., 2008. The effects of person-organization ethical fit on employee attraction and retention: Towards a testable explanatory model. *Journal of Business Ethics (JBE)* 78, 611–622. <https://doi.org/10.1007/s10551-007-9371-y>.
- Contractor, N., Monge, P., Leonardi, P.M., 2011. Multidimensional Networks and the Dynamics of Sociomateriality: Bringing Technology Inside the Network. *International Journal of Communication* 5, 682–720.
- Coombs, C., Stacey, P., Kawalek, P., Simeonova, B., Becker, J., Bergener, K., Carvalho, J., Fantinato, M., Garmann-Johnsen, N.F., Grimme, C., Stein, A., Trautmann, H., 2021. What is it about humanity that we can't give away to intelligent machines? A European perspective. *International Journal of Information Management (IJIM)* 58, 102311, 1–13. <https://doi.org/10.1016/j.ijinfomgt.2021.102311>.
- Cooren, F., 2004. Textual Agency: How texts do things in organizational settings. *Organization* 11, 373–393. <https://doi.org/10.1177/1350508404041998>.
- Cranefield, J., Doyle, C., 2022. “You have a few wordy spots”: Role-related impacts of working with an intelligent assistant, in: ECIS 2022 Research Paper. 30th European Conference on Information Systems (ECIS), Timisoara. 18.06.2022-24.06.2022, pp. 1–16.
- Daily, M., Oulasvirta, A., Rekimoto, J., 2017. Technology for Human Augmentation. *Computer* 50, 12–15. <https://doi.org/10.1109/MC.2017.39>.
- Dando, N., Swift, T., 2003. Transparency and Assurance: Minding the credibility gap. *Journal of Business Ethics (JBE)* 44, 195–200. <https://doi.org/10.1023/A:1023351816790>.
- Davenport, T.H., Kirby, J., 2016. Just how smart are smart machines? *MITSloan Management Review* 57, 21–25.
- Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., Ebel, P., 2019a. The future of Human-AI collaboration: A taxonomy of design knowledge for hybrid Intelligence Systems, in: 52nd Hawaii International Conference on System Sciences (HICSS). 52nd Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii. 08.01.2019-11.01.2019, pp. 1–10.
- Dellermann, D., Ebel, P., Söllner, M., Leimeister, J.M., 2019b. Hybrid Intelligence. *Business & Information Systems Engineering (BISE)* 61, 637–643. <https://doi.org/10.1007/s12599-019-00595-2>.
- Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H.,

- Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., Wright, R., 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>.
- Ehsan, U., Riedl, M.O., 2020. Human-centered explainable AI: Towards a reflective sociotechnical approach, in: 22nd International Conference on Human-Computer Interaction (HCI). 22nd International Conference on Human-Computer Interaction (HCI), Copenhagen, Dänemark. 19.07.2020-24.07.2020, pp. 1–19.
- Eitel-Porter, R., 2021. Beyond the Promise: Implementing Ethical AI. *AI Ethics* 1, 73–80. <https://doi.org/10.1007/s43681-020-00011-6>.
- Etzioni, A., Etzioni, O., 2017. Incorporating Ethics into Artificial Intelligence. *The Journal Of Ethics* 21, 403–418. <https://doi.org/10.1007/s10892-017-9252-2>.
- Faraj, S., Pachidi, S., Sayegh, K., 2018. Working and organizing in the age of the learning algorithm. *Information and Organization* 28, 62–70. <https://doi.org/10.1016/j.infoandorg.2018.02.005>.
- Faulkner, P., Runde, J., 2011. The social, the material, and the ontology of non-material technological objects. *European Group for Organizational Studies (EGOS)* -, 1–34.
- Figueroa-Armijos, M., Clark, B.B., Da Motta Veiga, S.P., 2022. Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and Social Influence. *Journal of Business Ethics (JBE)* -, 1–19. <https://doi.org/10.1007/s10551-022-05166-2>.
- Financial Times, 2016. 50 journals used in FT research rank. <https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0> (accessed 16 January 2022).
- Fjeld, J., Achten, N., Hilligoss, H.: Nagy, A., Srikumar, M., 2020. Principled Artificial Intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI, 39 pp. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420> (accessed 27 January 2022).
- Flathmann, C., Schelble, B.G., Zhang, R., McNeese, N.J., 2021. Modeling and guiding the creation of ethical Human-AI teams, in: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA*. 19.05.2021-21.05.2021. ACM, New York, NY, USA, pp. 469–479.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E., 2018. AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines* 28, 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.

- Fuegener, A., Grahl, J., Gupta, A., Ketter, W., 2021. Cognitive challenges in Human-AI collaboration: investigating the path towards productive delegation. *Information Systems Research (ISR)* 33, 678–696. <https://doi.org/10.1287/isre.2021.1079>.
- Galliers, R.D., Sutherland, A.R., 2009. The evolving Information Systems strategy, in: Galliers, R., Leidner, D.E. (Eds.), *Challenges and strategies in managing Information Systems*, vol. 3, 4th ed. Routledge, New York, pp. 1–31.
- Gandz, J., Hayes, N., 1988. Teaching Business Ethics. *Journal of Business Ethics (JBE)* 7, 657–669.
- Gasser, U., Almeida, V., 2017. A layered model for AI governance. *IEEE Internet Computing* 21, 58–62. <https://doi.org/10.1109/MIC.2017.4180835>.
- Giermindl, L.M., Strich, F., Christ, O., Leicht-Deobald, U., Redzepi, A., 2022. The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems (EJIS)* 31, 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>.
- Goodpaster, K.E., 1991. Business Ethics and stakeholder analysis. *Business Ethics Quarterly (BEQ)* 1, 53–73. <https://doi.org/10.2307/3857592>.
- Greene, T., Shmueli, G., Ray, S., forthcoming. Taking the Person Seriously: Ethically-aware IS Research in the Era of Reinforcement Learning-based Personalization. *JAIS Preprints (Journal of the Association for Information Systems)*, 1–35. <https://doi.org/10.17705/1jais.00491>.
- Greenwood, M.R., 2002. Ethics and HRM: A review and conceptual analysis. *Journal of Business Ethics (JBE)* 36, 216–278. <https://doi.org/10.1023/A:1014090411946>.
- Grønsund, T., Aanestad, M., 2020. Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems* 29, 101614, 1–16. <https://doi.org/10.1016/j.jsis.2020.101614>.
- Gupta, S., Kamboj, S., Bag, S., 2021. Role of risks in the development of responsible Artificial Intelligence in the digital healthcare domain. *Information Systems Frontiers (ISF)* -, 1–18. <https://doi.org/10.1007/s10796-021-10174-0>.
- Hamilton, C., Swart, W., Stokes, G.M., 2021. Developing a measure of social, ethical, and legal content for intelligent cognitive assistants. *Journal of Strategic Innovation and Sustainability* 16, 1–37. <https://doi.org/10.33423/jsis.v16i3>.
- Han, S., Kelly, E., Nikou, S., Svee, E.O., 2020. Reflections on Artificial Intelligence alignment with human values: A phenomenological perspective, in: 28th European Conference on Information Systems 2020 (ECIS). 28th European Conference on Information Systems (ECIS), Marrakech, Morocco. 15.06.2020-17.06.2020, pp. 1–17.
- Harzing, A.-W., 2020. Journal quality list, 24 pp. <https://harzing.com/>.
- Hasnas, J., 2013. Teaching Business Ethics: The principles approach. *Journal of Business Ethics Education (JBEE)* 10, 275–304. <https://doi.org/10.5840/jbee20131014>.
- Henderson, J.C., Venkatraman, H., 1999. Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal* 38, 472–484. <https://doi.org/10.1147/SJ.1999.5387096>.
- Heyder, T., Posegga, O., 2021. Extending the foundations of AI literacy, in: *Proceedings of the 42nd International Conference on Information Systems (ICIS)*, Austin, Texas.

- Hinsen, S., Hofmann, P., Jöhnk, J., Urbach, N., 2022. How can organizations design purposeful Human-AI interactions: A practical perspective from existing use cases and interviews, in: Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS). 55th Hawaii International Conference on System Sciences (HICSS-55), Maui, Hawaii (Virtual). 04.01.2022-07.01.2022. University of Hawaii at Manoa Hamilton Library, Honolulu, HI, pp. 1–11.
- Hoeddinghaus, M., Sondern, D., Hertel, G., 2021. The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior* 116, 1-57. <https://doi.org/10.1016/j.chb.2020.106635>.
- Hohenstein, J., Jung, M., 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106, 1-13. <https://doi.org/10.1016/j.chb.2019.106190>.
- Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., Wallach, H., 2019. Improving fairness in machine learning systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems (CHI), Glasgow, Scotland. 04.05.2019-09.05.2019. ACM, New York, NY, USA, pp. 1–16.
- Hosmer, L.T., 1987. Ethical Analysis and Human Resource Management. *Human Resource Management* 26, 313–330. <https://doi.org/10.1002/hrm.3930260302>.
- Islam, G., Greenwood, M., 2021. Reconnecting to the social in Business Ethics. *Journal of Business Ethics (JBE)* 170, 1–4. <https://doi.org/10.1007/s10551-021-04775-7>.
- Janssen, C.P., Donker, S.F., Brumby, D.P., Kun, A.L., 2019. History and future of human-automation interaction. *International Journal of Human-Computer Studies* 131, 99–107. <https://doi.org/10.1016/j.ijhcs.2019.05.006>.
- Jia, H., Wu, M., Sundar, S.S., 2022. Do we blame it on the machine? Task outcome and agency attribution in human-technology collaboration, in: Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS). 55th Hawaii International Conference on System Sciences (HICSS-55), Maui, Hawaii (Virtual). 04.01.2022-07.01.2022. University of Hawaii at Manoa Hamilton Library, Honolulu, HI, pp. 388–397.
- Jiang, J., Adam, M., Benlian, A., 2021. Algoactivistic practices in ridesharing - A topic modeling & grounded theory approach, in: Proceedings of the 29th European Conference on Information Systems (ECIS). 29th European Conference on Information Systems (ECIS), Marrakech, Morocco. 14.06.2021-16.06.2021, pp. 1–17.
- Jobin, A., Ienca, M., Vayena, E., 2019. Artificial Intelligence: The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>.
- Jones, C., Parker, M., Bos, R. ten, 2005. *For Business Ethics*. Routledge, London, UK, 223 pp.
- Jones, M., 2014. A matter of life and death: exploring conceptualizations of sociomateriality in the context of critical care. *MIS Quarterly (MISQ)* 38, 895-926. <https://doi.org/10.25300/MISQ/2014/38.3.12>.
- Kamar, E., 2016. Directions in hybrid intelligence: Complementing AI Systems with human intelligence, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI). Twenty-Fifth International Joint Conferences on Artificial Intelligence (IJCAI), New York, NY. 09.07.2016-15.07.2016, pp. 4070–4073.
- Kant, I., 1991. *The Metaphysics of Morals*. Cambridge University Press, Cambridge.

- Karimi, J., Konsynski, B.R., 2002. The Information Technology and management infrastructure strategy, in: Galliers, R.D., Leidner, D.E. (Eds.), *Strategic Information Management: Challenges and strategies in managing Information Systems*, 3rd Edition, 3rd ed. Routledge, London, UK, pp. 89–112.
- Kelley, S., 2022. Employee perceptions of the effective adoption of AI principles. *Journal of Business Ethics (JBE)* 178, 871–893. <https://doi.org/10.1007/s10551-022-05051-y>.
- Kitchenham, B., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering: Version 2.3, 65 pp. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf.
- Klumpp, M., 2018. Automation and Artificial Intelligence in business logistics systems: human reactions and collaboration requirements. *International Journal of Logistics Research and Applications* 21, 224–242. <https://doi.org/10.1080/13675567.2017.1384451>.
- Klumpp, M., Zijm, H., 2019. Logistics innovation and social sustainability: How to prevent an artificial divide in Human–Computer Interaction. *Journal of Business Logistics (JBL)* 40, 265–278. <https://doi.org/10.1111/jbl.12198>.
- Kolbjørnsrud, V., Amico, R., Thomas, R.J., 2017. Partnering with AI: how organizations can win over skeptical managers. *Strategy & Leadership (SL)* 45, 37–43. <https://doi.org/10.1108/SL-12-2016-0085>.
- Larsen, K.R., Hovorka, D.S., Dennis, A.R., West, J.D., 2019. Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles. *Journal of the Association for Information Systems (JAIS)* 20, 887–927. <https://doi.org/10.17705/1jais.00556>.
- Latour, B., 1992. Where are the missing masses? The sociology of a few mundane artifacts. *Shaping technology/building society: Studies in sociotechnical change* 1, 225–258.
- Latour, B., 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, Oxford, New York, 301 pp.
- Law, J., 1992. Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice* 5, 379–393. <https://doi.org/10.1007/BF01059830>.
- Lee, M.K., 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society (BD&S)* 5, 1-16. <https://doi.org/10.1177/2053951718756684>.
- Lee, M.K., Kusbit, D., Metsky, E., Dabbish, L., 2015. Working with machines, in: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15: CHI Conference on Human Factors in Computing Systems (CHI)*, Seoul Republic of Korea. 18.04.2015-23.04.2015. ACM, New York, NY, USA, pp. 1603–1612.
- Leonardi, P.M., 2011. When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly (MISQ)* 35, 147–167. <https://doi.org/10.2307/23043493>.
- Leonardi, P.M., 2012. Materiality, Sociomateriality, and Socio-Technical Systems: What do these terms mean? How are they related? Do we need them?, in: Leonardi, P.M., Nardi, B.A., Kallinikos, J. (Eds.), *Materiality and organizing: Social Interaction in a technological world*. Oxford University Press, Oxford, UK, pp. 25–48.
- Leslie, D., 2019. *Understanding Artificial Intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute, 97 pp.

- Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20: CHI Conference on Human Factors in Computing Systems (CHI), Honolulu HI USA. 25.04.2020-30.04.2020. ACM, New York, NY, USA, pp. 1–15.
- Long, D., Magerko, B., 2020. What is AI literacy? Competencies and design considerations, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20: CHI Conference on Human Factors in Computing Systems (CHI), Honolulu HI USA. 25.04.2020-30.04.2020. ACM, New York, NY, USA, pp. 1–16.
- Lowry, P.B., Gaskin, J., Humpherys, S.L., Moody, G.D., Galletta, D.F., Barlow, J.B., Wilson, D.W., 2013. Evaluating journal quality and the association for information systems senior scholars' journal basket via bibliometric measures: Do expert journal assessments add value? *MIS Quarterly (MISQ)* 37, 993–1012. <https://doi.org/10.25300/MISQ/2013/37.4.01>.
- MacInnis, D.J., 2011. A framework for conceptual contributions in marketing. *Journal of Marketing (JM)* 75, 136–154. <https://doi.org/10.1509/jmkg.75.4.136>.
- MacKenzie, S.B., 2003. The Dangers of Poor Construct Conceptualization. *Journal of the Academy of Marketing Science* 31, 323–326. <https://doi.org/10.1177/0092070303031003011>.
- Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H., 2020. Co-Designing checklists to understand organizational challenges and opportunities around fairness in AI, in: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI). CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI), Honolulu, USA. 25.04.2020-30.04.2020. Association for Computing Machinery, New York, NY, USA, pp. 1–14.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., Söllner, M., 2019. AI-Based digital assistants. *Business & Information Systems Engineering (BISE)* 61, 535–544. <https://doi.org/10.1007/s12599-019-00600-8>.
- Malik, N., Tripathi, S.N., Kar, A.K., Gupta, S., 2021. Impact of artificial intelligence on employees working in industry 4.0 led organizations. *International Journal of Manpower (IJM)* 43, 334–354. <https://doi.org/10.1108/IJM-03-2021-0173>.
- Marabelli, M., Newell, S., Handunge, V., 2021. The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. *Journal of Strategic Information Systems* 30, 101683. <https://doi.org/10.1016/j.jsis.2021.101683>.
- Markus, M.L., Rowe, F., 2021. Guest editorial: Theories of digital transformation: A progress report. *Journal of the Association for Information Systems (JAIS)* 22, 273–280. <https://doi.org/10.17705/1jais.00661>.
- Markus, M.L., Rowe, F., 2023. The Digital Transformation Conundrum: Labels, Definitions, Phenomena, and Theories. *JAIS* 24, 328–335. <https://doi.org/10.17705/1jais.00809>.
- Mayer, A.S., Haimlerl, A., Strich, F., Fiedler, M., 2021. How corporations encourage the implementation of AI ethics, in: ECIS 2021 Research Papers. 29th European Conference on Information Systems (ECIS), Marrakech, Morocco. 14.06.2021-16.06.2021, 1-15.

- Mayer, A.-S., Strich, F., Fiedler, M., 2020. Unintended consequences of introducing AI systems for decision making. *MIS Quarterly Executive (MISQE)* 19, 239–257. <https://doi.org/10.17705/2msqe.00036>.
- Mayring, P., 2014. *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt.
- McHugh, M.L., 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 276–282. <https://doi.org/10.1201/b16720>.
- Mingers, J., Harzing, A.W., 2007. Ranking journals in business and management: A statistical analysis of the Harzing data set. *European Journal of Information Systems (EJIS)* 16, 303–316. <https://doi.org/10.1057/palgrave.ejis.3000696>.
- Mirbabaie, M., Brünker, F., Möllmann Frick, N., Stieglitz, S., 2022. The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets* 32, 73–99. <https://doi.org/10.1007/s12525-021-00496-x>.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. *Nature* 518, 529–533. <https://doi.org/10.1038/nature14236>.
- Mollick, E., 2022. ChatGPT is a tipping point for AI. *Harvard Business Review*.
- Mutch, A., 2013. Sociomateriality — Taking the wrong turning? *Information and Organization* 23, 28–40. <https://doi.org/10.1016/j.infoandorg.2013.02.001>.
- Ng, D., Leung, J., Chu, S., Qiao, M.S., 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2, 1–11. <https://doi.org/10.1016/j.caeai.2021.100041>.
- Norman, D., 2017. Design, business models, and human-technology teamwork: Everything as a service. *Research-Technology Management (RTM)* 60, 26–30. <https://doi.org/10.1080/08956308.2017.1255051>.
- Orlikowski, W.J., 2007. Sociomaterial practices: Exploring technology at work. *Organization Studies (OS)* 28, 1435–1448. <https://doi.org/10.1177/0170840607081138>.
- Orlikowski, W.J., 2010. The sociomateriality of organisational life: considering technology in management research. *Cambridge Journal of Economics (CJE)* 34, 125–141. <https://doi.org/10.1093/cje/bep058>.
- Orlikowski, W.J., Scott, S.V., 2008. 10 Sociomateriality: Challenging the separation of technology, work and organization. *The Academy of Management Annals (Annals)* 2, 433–474. <https://doi.org/10.1080/19416520802211644>.
- O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., Holzinger, K., Holzinger, A., Sajid, M.I., Ashrafi, H., 2019. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *International Journal of Medical Robotics and Computer Assisted Surgery* 15, e1968. <https://doi.org/10.1002/rcs.1968>.
- Paré, G., Trudel, M.-C., Jaana, M., Kitsiou, S., 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52, 183–199. <https://doi.org/10.1016/j.im.2014.08.008>.

- Park, H., Ahn, D., Hosanagar, K., Lee, J., 2021. Human-AI Interaction in Human Resource Management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens, in: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21: CHI Conference on Human Factors in Computing Systems (CHI 2021), Yokohama, Japan. 08.05.2021-13.05.2021. Association for Computing Machinery, New York, New York, pp. 1–15.
- Parker, M., 2002. *Against Management: Organization in the age of managerialism*. Oxford University Press, Oxford, UK.
- Pavlik, J.V., 2023. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator* 78, 84–93. <https://doi.org/10.1177/10776958221149577>.
- Payne, S.L., Wayland, R.F., 1999. Ethical obligation and diverse values assumptions in HRM. *International Journal of Manpower (IJM)* 20, 297–308. <https://doi.org/10.1108/01437729910286083>.
- Peppard, J., Ward, J., 2004. Beyond strategic information systems: towards an IS capability. *The Journal of Strategic Information Systems* 13, 167–194. <https://doi.org/10.1016/j.jsis.2004.02.002>.
- Polyviou, A., Zamani, E.D., 2022. Are we Nearly There Yet? A Desires & Realities Framework for Europe's AI Strategy. *Information Systems Frontiers*, 1–17. <https://doi.org/10.1007/s10796-022-10285-2>.
- Pregenzer, M., Remus, U., Wiener, M., 2021. Algorithms in the driver's seat: explaining workers' reactions to algorithmic control, in: *Proceedings of the 29th European Conference on Information Systems (ECIS)*. 29th European Conference on Information Systems (ECIS), Marrakech, Morocco. 14.06.2021-16.06.2021, pp. 1–15.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training. Preprint.
- Rai, A., Constantinides, P., Sarker, S., 2019. Next generation digital platforms: Toward Human-AI hybrids. *MIS Quarterly (MISQ)* 43, 3-9.
- Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., Farooq, A., 2019. Human augmentation: Past, present and future. *International Journal of Human-Computer Studies* 131, 131–143. <https://doi.org/10.1016/j.ijhcs.2019.05.008>.
- Raisch, S., Krakowski, S., 2020. Artificial Intelligence and Management: The automation-augmentation paradox. *Academy of Management Review* 46, 192–210. <https://doi.org/10.5465/2018.0072>.
- Rajaobelina, L., Prom Tep, S., Arcand, M., Ricard, L., 2021. Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology & Marketing* 38, 2339–2356. <https://doi.org/10.1002/mar.21548>.
- Rakowski, R., Polak, P., Kowalikova, P., 2021. Ethical aspects of the impact of AI: The status of humans in the era of Artificial Intelligence. *Society* 58, 196–203. <https://doi.org/10.1007/s12115-021-00586-8>.
- Reckwitz, A., 2002. Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory (EJST)* 5, 243–263. <https://doi.org/10.1177/1368431002005002005>.
- Rezaee, Z., 2008. *Corporate governance and ethics*. John Wiley & Sons, Boston, MA.

- Rieder, T.N., Hutler, B., Mathews, D., 2020. Artificial Intelligence in service of human needs: Pragmatic first steps toward an ethics for semi-autonomous agents. *AJOB Neuroscience* 11, 120–127. <https://doi.org/10.1080/21507740.2020.1740354>.
- Ritson, N., 2008. *Strategic Management*, 2nd ed. bookboon.com, 117 pp.
- Rivard, S., 2021. Theory building is neither an art nor a science: It is a craft. *Journal of Information Technology (JIT)* 36, 316–328. <https://doi.org/10.1177/0268396220911938>.
- Rix, J., Hess, T., 2022. Hello, Mate! Insights from the field on leveraging machine teammates in organizations, in: *Proceedings of the 26th Pacific Asia Conference on Information Systems (PACIS)*. 26th Pacific Asia Conference on Information Systems (PACIS), Virtual Event / Taipei, Taiwan / Sydney, Australia. 05.07.2022–09.07.2022.
- Saupé, A., Mutlu, B., 2015. The social impact of a robot co-worker in industrial settings, in: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15: 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI), Seoul Republic of Korea. 18.04.2015–23.04.2015. ACM, New York, NY, USA, pp. 3613–3622.
- Scherer, M.U., 2016. Regulating Artificial Intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology* 29, 354–398. <https://doi.org/10.2139/ssrn.2609777>.
- Schwartz, M.S., Saiia, D., 2012. Should firms go beyond profits? Milton Friedman versus broad CSR. *Business and Society Review* 117, 1–13. <https://doi.org/10.1111/j.1467-8594.2011.00397.x>.
- Schwarz, A., Mehta, M., Johnson, N., Chin, W.W., 2007. Understanding Frameworks and Reviews: A Commentary to Assist us in Moving Our Field Forward by Analyzing Our Past. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 38, 29–50.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., Freitas, N. de, 2016. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* 104, 148–175. <https://doi.org/10.1109/JPROC.2015.2494218>.
- Shin, D., Park, Y.J., 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98, 277–284. <https://doi.org/10.1016/j.chb.2019.04.019>.
- Shneiderman, B., 2020. Human-centered Artificial Intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 109–124. <https://doi.org/10.17705/1thci.00131>.
- Siau, K., Wang, W., 2020. Artificial Intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management (JDM)* 31, 74–87. <https://doi.org/10.4018/JDM.2020040105>.
- Someh, I., Wixom, B.H., Beath, C.M., Zutavern, A., 2022. Building an Artificial Intelligence explanation capability. *MIS Quarterly Executive (MISQE)* 21, 5, 143–163. <https://doi.org/10.17705/2msqe.00063>.
- Somogyi, E.K., Galliers, R.D., 2009. Developments in the application of Information Technology in business, in: Galliers, R., Leidner, D.E. (Eds.), *Challenges and strategies in managing Information Systems*, 4th ed. Routledge, New York, pp. 3–22.
- Stevens, B., 1999. Communicating ethical values: A study of employee perceptions. *Journal of Business Ethics (JBE)* 20, 113–120.

- Strich, F., Mayer, A.-S., Fiedler, M., 2021. What do I do in a world of Artificial Intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems (JAIS)* 22, 304–324. <https://doi.org/10.17705/1jais.00663>.
- Te'eni, D., Avital, M., Hevner, A., Schoop, M., Schwartz, D.G., 2019. It takes two to tango: Choreographing the interactions between Human and Artificial Intelligence, in: *27th European Conference on Information Systems: Information Systems for a Sharing Society (ECIS)*. 27th European Conference on Information Systems: Information Systems for a Sharing Society (ECIS), Stockholm, Sweden. 08.06.2019-14.06.2019, -.
- Templier, M., Paré, G., 2018. Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems (EJIS)* 27, 503–550. <https://doi.org/10.1080/0960085X.2017.1398880>.
- Thompson, M., 2012. People, practice, and technology: Restoring Giddens' broader philosophy to the study of information systems. *Information and Organization* 22, 3, 188–207. <https://doi.org/10.1016/j.infoandorg.2012.04.001>.
- VHB, 2011. JOURQUAL-3 Gesamtranking. <https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/tabellen-zum-download> (accessed 27 January 2022).
- Vial, G., 2019. Understanding digital transformation: A review and a research agenda. *Journal of Strategic Information Systems* 28, 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>.
- Wang, Y., Xiong, M., Olya, H., 2020. Toward an understanding of responsible Artificial Intelligence practices, in: *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*. 53rd Hawaii International Conference on System Sciences 2020 (HICSS-53), Maui, Hawaii. 07.01.2020-10.01.2020, pp. 4962–4971.
- Watson, G.J., Desouza, K.C., Ribiere, V.M., Lindič, J., 2021. Will AI ever sit at the C-suite table? The future of senior leadership. *Business Horizons* 64, 465–474. <https://doi.org/10.1016/j.bushor.2021.02.011>.
- Weber, R., 2012. Evaluating and Developing Theories in the Information Systems Discipline. *Journal of the Association for Information Systems (JAIS)* 13, 1–30. <https://doi.org/10.17705/1jais.00284>.
- Whittlestone, J., Nyrop, R., Alexandrova, A., 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. <https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf> (accessed 27 January 2022).
- Wilson, J., Daugherty, P.R., 2018. Collaborative Intelligence: Humans and AI are joining forces. *Harvard Business Review* 96, 114–123.
- Winstanley, D., Woodall, J., Heery, E., 1996. Business ethics and human resource management: Themes and issues. *Personnel Review* 25, 5–12. <https://doi.org/10.1108/00483489610148491>.
- Zhang, X., Zhao, Y., Tang, X., Zhu, H., Xiong, H., 2020. Developing fairness rules for talent intelligence management system, in: *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*. 53rd Hawaii International Conference on System Sciences (HICSS-53), Maui, Hawaii. 07.01.2020-10.01.2020, pp. 5882–5891.

Paper VII

Hammerschmidt, T. (2025). *Navigating the Nexus of Ethical Standards and Moral Values*. Ethics and Information Technology (EIT), 27(17), <https://doi.org/10.1007/s10676-025-09826-5>.

Navigating the Nexus of Ethical Standards and Moral Values

Full Research Article – Accepted Version

Abstract. This study examines how ethical standards established by stakeholders such as developers and policymakers provide top-down guidance aligned with deontological ethics or utilitarian goals. It also highlights a complementary bottom-up approach, rooted in virtue ethics, in which individuals engage in ethical deliberations shaped by their moral values. Both approaches have limitations, and, at times, ethical standards can clash with moral values, thus blurring lines of responsibilities. Deontological principles may offer a structured framework, but often lack adaptability to diverse cultural contexts; bottom-up approaches foster intrinsic moral intentions, but universal applicability may be challenging, thus raising moral dilemmas. Through a theoretical literature review, this study explains how different ontological and normative ethical perceptions lead to moral dilemmas in various AI application scenarios (e.g., algorithmically managed platforms, crime detection systems, medical AI assistants). It addresses top-down and bottom-up approaches that may help account for moral dilemmas ethically. The study discusses the balance between top-down regulatory frameworks and bottom-up community-driven ethics to navigate the complex ethical landscape of AI applications, whose increasing capabilities alter expectations of AI's agency and morality. This study calls for holistic and multi-objective ethical frameworks that incorporate diverse normative ethical perspectives and recognizes context-specific ontologies throughout the AI lifecycle. It emphasizes a nuanced and context-specific combination of top-down standards (e.g., regulatory oversight, clear guidelines) and bottom-up fostering of moral values (e.g., by improving ethical knowledge). This tailored and ongoing reflection of ethical standards and moral values accounts for an ethical development, deployment, and utilization of AI technologies.

Keywords: Literature Review, AI Ethics, AI Governance, Normative Ethics

1. Introduction

Artificial intelligence (AI) technologies have advanced from simple rule-based systems into complex models that inform, predict, and guide decisions (Berente et al., 2019; Dwivedi et al., 2021). These changes in AI capabilities have shifted the traditional power dynamics between humans and machines, moving AI from a subordinated role to one of autonomy (Giermindl et al., 2022; Maedche et al., 2019). With AI increasingly integrated into society and business, it is essential to consider ethical standards such as fairness, accountability, and transparency to ensure that AI is designed, deployed, and used responsibly (Aslan et al., 2022; Mirbabaie et al., 2022). While fairness, accountability, and transparency are often framed as values, Winfield (2019) suggests it is useful to approach them through the lens of “ethical standards,” as these standards

encompass shared norms and formalized principles of core ethical values developed through consensus among various stakeholders—developers, organizational managers, governmental representatives, market parties, and so on—“to, at best remove, hopefully reduce, or at very least highlight the potential for unethical impacts of consequences” of using AI (Winfield, 2019, p. 46). Even though the diversity of moral values among stakeholders poses challenges to establishing clear ethical standards (e.g., Badea & Artus, 2022; Telkamp & Anderson, 2022), these standards provide guidance within specific communities to help determine ethically responsible behaviors in contexts involving AI (Siau & Wang, 2020). This top-down approach to ethics, closely related to deontological ethics, relies on principles rooted in duty, norm compliance, and obligation (Millar, 2016; Swanepoel, 2021; Unver, 2023).

Conversely, a complementary bottom-up approach allows individuals to engage in ethical deliberation informed by personal experiences and societal influences (Roberts & Montoya, 2022). This bottom-up approach draws on virtue ethics, where an individual’s moral convictions are shaped by cultural, religious, societal, and philosophical opinions and are cultivated through intrinsic virtues such as justice, courage, temperance, and prudence (Graff, 2024; Neubert & Montañez, 2020), enabling individuals to either conform with or diverge from established standards (Bogosian, 2017).

Both the top-down implementation of ethical standards (deontology) and bottom-up acquisition of moral values (virtue ethics) have inherent limitations. While bottom-up learning risks unintentionally generalizing examples in undesired ways, top-down standards, because they depend on stakeholder consensus within specific communities, may fail to address complex situations (Constantinescu et al., 2021; Graff, 2024; Meijer et al., 2023). Additionally, the unsupervised learning capacities of AI result in AI increasingly adapts based on user data and feedback, reflecting user values and virtues, besides adhering to given ethical principles by design (Etzioni & Etzioni, 2017; Greene et al., 2023; Peters et al., 2020). This user feedback-based ethical sense-making contrasts with traditional, top-down ethical standards, underscoring the need for flexible ethical frameworks that account for diverse moral inputs (Siau & Wang, 2020).

The IEEE 7000 standards initiative, which brought together experts from various cultures and professional backgrounds, emphasizes the importance of integrating local ethical traditions when establishing AI ethics standards, reflecting a variety of normative ethical approaches, including deontology, virtue ethics, and utilitarianism (Spiekermann, 2021). The alignment of ethical standards and individual moral values are increasingly seen as essential for mitigating moral dilemmas in AI use (e.g., Bankins & Formosa, 2023; Buhmann & Fieseler, 2023; Hagendorff, 2020; Rodgers et al., 2023). Absent alignment, ethical standards may not be socially accepted and may give rise to conflicts in value (Buruk et al., 2020; Córdova & Vicari, 2022). Ethical standards can serve as guiding principles for determining responsible behavior and making subsequent judgments, but they also may clash with individuals’ behavior with respect to moral blame for wrongdoing (McDonald & Pan, 2020). Some scholars argue that the primary object should not revolve solely around alignment, as this approach imposes limitations by setting a minimum threshold for acceptable ethical standards, potentially

lagging behind higher moral values of individuals (Constantinescu et al., 2021; Meijer et al., 2023).

Hence, there is a need to increase our understanding of the interplay between ethical standards and moral values within human-AI interaction by developing ethical frameworks that adhere to several normative ethical theories (Wickson & Forsberg, 2015). While there has been some valuable research on different normative ethical theories in the context of human-AI interaction (e.g., Heyder et al., 2023; Serafimova, 2020; Telkamp & Anderson, 2022), further investigations are required for a deeper analysis of how ethical standards can either strengthen or impair moral values and vice versa, and how this may lead to moral dilemmas in organizations. Particularly, sociotechnical systems (STS) can be a useful framework as it views organizations as integrated systems where social entities (like people, roles, and culture) and technical elements (like tools, processes, and technologies) interact closely and shape one another, such as moral values shape ethical standards and vice versa. Based on a systematic literature review, this work aims to answer the following research question is: *How can normative ethical theories help explain the relationship between ethical standards and moral values in human interaction with AI technologies?*¹

This study offers a comprehensive analysis of how ethical standards can either bolster or undermine moral values within organizations and, conversely, how these dynamics can impact the ethical design, deployment and outcomes of AI technologies. In doing so, this work explains how and why moral dilemmas may arise when humans interact with AI technologies according to their different ontological and normative ethical assumptions and further examines the underlying top-down and bottom-up approaches that may help account for those conflicts.

2. Integrating Normative Ethics in Sociotechnical Systems: The Interplay of Moral Values and Ethical Standards

This section describes how different normative ethical approaches shape human-AI interactions in organizations as STS. The sociotechnical perspective highlights interdependencies among human users, AI technologies, and broader organizational entities (e.g., stakeholders, societal, norms, and regulation). Building upon Leonardi's (2012) conceptualization of STS, the interaction between AI and human users can be understood as *sociomaterial entanglement* of the *material entity* (AI technology) and the *social entity* (human user). The *social subsystem* consists of additional organizational stakeholders, processes, norms, and values that further shape human-AI interactions, influencing perceptions of ethical or unethical behavior (Ceccez-Kecmanovic et al., 2014). The three main ethical approaches, deontology, virtue ethics,

¹ This study uses the term "AI technologies" more broadly to refer to any technological artifact with AI whose action may have normatively relevant consequences. This includes specific types of AI technologies, such as autonomous vehicles. Although not all findings about one technology apply to others, this study aggregates findings regarding various AI technologies being applied in different domains and ethical contexts, as an overly focused discussion could inadvertently omit crucial aspects of responsible AI usage and outcomes.

and utilitarianism, may explain different assumptions regarding ethical or non-ethical behavior (Bankins & Formosa, 2023; Bilal et al., 2021). The following sections describe how these ethical approaches manifest within STS.

2.1. How Ontological Frameworks Impact Normative Ethical Perceptions

Ontology, the philosophical study of being, provides frameworks to understand the fundamental nature, properties, and relationships of entities within an STS, including *material* and *social entities*, their agency, their actions, and the outcomes of the *sociomaterial entanglement* (Orlikowski, 2010). The environment—with its actors, objects, and artifacts (i.e., *social subsystem*)—creates different realities regarding ethical AI design, deployment, and usage and, thus, determining the ontology (Cunneen et al., 2019). This, in turn, influences our understanding of what ethical standards and moral values might be deemed appropriate.

Relational and substantial ontologies are typical of the ontological frameworks that have been discussed within the sociomateriality and STS literature (Cecez-Kecmanovic et al., 2014;). Relational ontology posits that entities do not have intrinsic properties independent of their relations with other entities (Barad, 2003; Latour, 1992), whereas substantial ontology posits that entities have an independent and intrinsic existence that is characterized by stable properties (Faulkner & Runde, 2010; Mutch, 2013). In the context of human-AI interaction, relational ontology suggests that the nature and behavior of both humans and AI systems are shaped by their dynamics, and that their roles are co-constructed through the *sociomaterial entanglement* (Heyder et al., 2023). In comparison, substantial ontology would treat AI technologies and humans as distinct entities with inherent characteristics that do not change (Heyder et al., 2023).

Ontological frameworks lead to varying ethical priorities among stakeholders within an STS, such as managers, developers, politicians, and users. They may have conflicting views regarding necessary ethical standards given their heterogeneous underlying ontological frameworks. These different ontological frameworks can lead to varied preferences for ethically designed AI technologies (Kieslich et al., 2022), thereby shaping perceptions about accountability as a societal normative obligation for dealing with unintended outcomes of human-AI interaction (Horneber & Laumer, 2023; Novelli et al., 2023). Cunneen et al. (2019) argue that the analysis of ontology must precede ethical analysis. In Section 4, this study outlines the different ontological frameworks that guide normative ethical perceptions in the reviewed literature.

2.2. How Moral Values Shape Agency Within the Entanglement

Within STS, human agency (*social entity*), characterized by intentions and behavioral maxims, governs how individuals respond to technologies (Leonardi, 2012). This agency is shaped by people's ethical ideology, moral values, and responsibility perceptions, which are grounded in virtue ethics (Constantinescu et al., 2021). Consequently, people can decide to adhere to their moral values (virtue ethics), pursue their personal goals (utilitarianism), or comply with regulatory frameworks and ethical standards (deontology) (Heyder et al., 2023; Swanepoel, 2021). Human agency,

thereby, reflects the interplay between internal moral values and external normative influences.

AI, in contrast, lacks intrinsic moral considerations. The agency of the AI technology (*material entity*) can be described as “the way objects act when humans provoke it” (Leonardi, 2012, p. 37), influenced primarily by given design principles and pre-defined parameters of AI models (Shneiderman, 2020). While some scholars argue that this can be mainly linked to deontology (e.g., Heyder et al., 2023; Hooker & Kim, 2018), Yu et al. (2021) contend that the adoption of AI’s agency is based on hedonic principles (utilitarianism). This implies that AI is programmed to maximize outputs, resulting in people delegating more tasks to AI technologies with significant social consequences. Others argue that AI can, to a limited extent, reflect human moral behavior in line with people’s virtue ethics, extending beyond ethical standards; this results from AI technologies’ abilities to learn from human data inputs (e.g., Etzioni & Etzioni, 2017; Moor, 2006). Even though AI can process data that involves moral considerations of users, it does not possess an understanding of moral values similar to human agency.

Thus, people have the free will to decide according to their moral values as moral agents (Coeckelbergh, 2020), but AI lacks moral agency—“a computationally correct answer is not a moral outcome” (Puri, 2020, p. 52). In Section 4, this study analyses the impact of moral values on AI and human agency in the reviewed literature.

2.3. How Ethical Standards from the Social Subsystem Guide Entities

With respect to deontology, ethical standards are established by various stakeholders—such as developers, managers, and policymakers—within what Leonardi (2012) terms the *social subsystem*. Ethical standards serve as design guidelines for AI technologies to perform certain tasks within legal frameworks; they can also help with people’s perceptions of what is “right” or “wrong,” thus guiding their decisions and leading to more ethical AI use (Behdadi & Munthe, 2020; Champagne & Tonkens, 2023).

Guiding human users is required because moral agency alone does not ensure ethical behavior (Reynolds & Ceranic, 2007). People may misuse AI for personal gain or encounter moral dilemmas due to conflicting values (Lee, 2018; Mayer et al., 2020). In such cases, the presence of moral agency alone is insufficient to prevent unethical behavior, and ethical standards become essential. Robust ethical standards are required to ensure a responsible *sociomaterial entanglement*. For human users, ethical standards can cultivate moral awareness and promote ethical behavior. For AI technologies, ethical standards involve testing for biases, ensuring transparency, and establishing mechanisms to respect privacy and attribute accountability to those responsible for AI-driven actions. Some researchers, therefore, refer to an artificial moral agency that differs from traditional human agency and that is further limited and guided by human moral agency (e.g., Behdadi & Munthe, 2020; Tigard, 2021). In Section 4, this study analyses the impact of ethical standards on AI and human agency in the reviewed literature.

2.4. How Top-down and Bottom-up Approaches Interlock

A holistic approach to ethical human-AI interaction requires combining top-down establishment of ethical standards and norms (e.g., laws and governance guidelines) with bottom-up integration of societal values reflecting diverse moral communities.

AI adhering strictly to pre-defined ethical standards (deontology) may be limited by the ethical understanding of individual developers (virtue ethics of a single community) and could cause more harm than if an AI were to learn ethical reasoning from the collective input of diverse human users with varying moral perceptions (collective virtue ethics). Top-down ethical standards, like regulations or utilitarian principles for AI design, provide a necessary starting point but establish limited legal boundaries (Hagendorff, 2020; Hawkins & Mittelstadt, 2023). Incorporating Aristotle's concept of dianoetic virtues (i.e., learnable intellectual and moral excellence) and practical wisdom into AI can help to consider specific situational and circumstantial elements that influence moral behavior (i.e., the context of action) (Constantinescu et al., 2021). This method allows AI technologies to evolve and adapt based on real-world inputs and societal feedback, ensuring that the ethical reasoning embedded within these systems reflects a broader spectrum of moral values and perspectives (Cook, 2023; Etzioni & Etzioni, 2017; Saariluoma & Leikas, 2020).

Integrating top-down and bottom-up approaches can lead to human-AI interactions that are more ethical than if developers were to embed single principles in AI design (Cook, 2023; Saariluoma & Leikas, 2020). Such dual approaches enable a balance between standard ethical guidelines and adaptable, context-sensitive AI behavior that reflects diverse moral perspectives within the *sociomaterial entanglement*. This, in turn, may influence future AI regulations, thereby impacting the *social subsystem* with its top-down, deontological approaches. It is crucial to recognize different moral communities with unique ethical identities—not only for guiding human users but also for guiding AI design (Bryson, 2018; Champagne & Tonkens, 2023; Horváth, 2022).

In Section 4, this study shows how top-down regulatory frameworks are combined with bottom-up moral values in the literature to handle complex moral dilemmas.

3. Method

This study uses a theoretical literature to identify, describe, and reconfigure relevant findings into a novel framework (Paré et al., 2015). The initial search, conducted in October 2023 and updated in March 2024, followed a systematic procedure recommended by Templier and Paré (2018).

Literature Search: This study assessed five databases that index scientific publications relevant to our review to ensure broad coverage of literature: ACM Digital Library, AIS Electronic Library, Business Source Complete, Science Direct, and Web of Science. The search query for title, abstracts, and keywords—if available—was: (“human-AI”) AND (“ethic*” OR “moral*”) AND (“standards” OR “principles” OR “norms” OR “guidelines” OR “values” OR “conflict” or “dilemma”). The final search yielded 2,161 overall hits and 2,090 unique hits after the elimination of 71 duplicates.

Literature Selection and Format Screening: This study conducted a format screening based on format exclusion criteria as outlined in Appendix I. Whenever possible, these criteria were applied using database filter options. As a result, 653 articles were removed during format screening, leaving 1,437 articles.

Quality Assessment: Quality was assessed based on articles having been published in outlets within the Harzing Ranking (Harzing, 2005; Mingers & Harzing, 2007) and the VHB JOURQUAL-3 Ranking (VHB, 2015). Given the focus on the two domains ethics and AI, this study further included articles published in AI- or ethics-related outlets, even if they were not part of the VHB JOURQUAL-3 Ranking (see Appendix II). Consequently, 1,009 articles were removed, leaving 428 articles.

Relevance Screening and Data Extraction: Data were extracted only from relevant articles. The relevance screening was conducted through a full-text assessment based on three criteria (see Appendix III). That eliminated 402 articles, leaving 26 articles in the final sample (see Appendix IV). The final sample was imported into the software MAXQDA for qualitative analysis.

Data Analysis: This study developed a coding scheme for qualitative analysis, following Mayring (2014). Deductive coding following a coding scheme with definitions for codes, coding rules, and anchored examples was used to structure the content of relevant articles along the constructs of the theoretical background (Mayring, 2014, pp. 95–98). This was used to code ontological frameworks, normative ethical theories, and top-down and bottom-up approaches linked to ethical standards and moral values. In addition, inductive coding was used to code agency expectations and possible pathways to balance bottom-up and top-down approaches for an ethical AI design, deployment, and utilization. A snapshot of coding is given in Appendix V.

4. Results

The analysis reveals that different expectations about embedding moral values in AI stems from distinct ontological frameworks. Section 4.1 examines these frameworks leading to different moral agency expectations. Section 4.2 analyzes the reasons behind the occurrence of moral dilemmas. Section 4.3 delineates seven real-world examples with context-specific analyses. Section 4.4 presents pathways for balancing top-down approaches with bottom-up approaches.

4.1. Ontological Frameworks Form Moral Agency Expectations

The reviewed literature reveals that different ontological frameworks shape perceptions of ethics in AI.

AI agency is determined by STS entities: Within the analyzed literature, it becomes evident that AI's agency is described primarily through its level of abilities (e.g., cognitive abilities) and performance (e.g., accuracy). For instance, Fernandes et al. (2020) and Ibáñez and Olmeda (2022) define agency via the performance level of AI technologies compared to humans, noting that people can make errors due to tiredness and that AI can analyze larger quantities of data in less time. However, they

note that AI lacks emotional intelligence, as it lacks the personal and cultural experiences essential for adapting to unforeseen situations. Thus, while human agency is based on inherent values (substantial ontology), AI's agency is determined externally by design guidelines, data, and functions (relational ontology) (Cunneen et al., 2019).

Human agency affected by AI: As Endacott and Leonardi (2022, p. 465) illustrate, “AI [communication technologies] that are configured to operate autonomously not only make decisions about communication, but act on them. For example, they do not just determine whom to invite to a particular meeting, when to invite them, how long the meeting should be, and through what communication channel the meeting should occur, but they can also proceed to invite the attendees. When AI [communication technologies] act autonomously the principal loses a certain degree of control over the communication environment.” Hence, advances in AI technologies impact the sociomaterial entanglement and, consequently, human agency, which can be linked to relational ontology when considering how human agency is affected by AI.

Misleading perceptions and errors within normative spaces: Misinterpretations of AI's agency, stemming from differing ontological views, can lead to normative errors. Badea and Artus (2022) identify two types of mistakes: *instrumental* (fact-based errors due to flawed data or processes) and *intention* mistakes (ethical errors tied to developers' assumptions). Hence, the agency of AI is fundamentally shaped by the distinct nature of humans as social beings (relational ontology), forming their perceptions of humans compared to AI's unique characteristics and capabilities. Zhang et al. (2021) illustrate how participants perceive AI as inherently utilitarian, believing AI maximizes collective welfare, which illustrates *intention* mistakes. Similarly, Giroux et al. (2022) show that anthropomorphizing AI may lead people to attribute moral agency to it mistakenly, resulting in *intention* mistakes. In contrast, *instrumental* mistakes in AI arise when biased datasets lead to skewed outcomes, causing moral concerns despite perceptions that AI operates neutrally (Hong et al., 2020; Ratoff, 2021). People fail to differentiate between AI and human agency, leading to normative errors.

Different ontological frameworks determine AI's morality and legal status: While people can act according to common-sense morality—either through learning from environmental interactions or their own pro-moral motivation (substantial ontology)—AI's abilities to acquire pro-moral motivations may be limited to its programming to adapt to users' responses (relational ontology) (Ratoff, 2021). Ratoff (2021) argues that the continuous cognitive process of hypotheses generation, prediction, and error correction may enable AI technologies to align with human values. Thereby, AI replicates, to some extent, ethical reasoning via predictive processing models, potentially granting it some form of moral agency, which may endow these technologies with a legal status (Dahiyat, 2021). This would represent a substantial ontological framework, which raises the critical question posed by Dahiyat (2021, p. 67): “At what point or degree of autonomy and sophistication would an agent acquire personality?” The legal status of AI technologies and responsibility for their actions and decisions may depend on the degree of agency attributed to them as material entities in contrast to human ones (substantial vs. relational ontology).

4.2. Why Moral Dilemmas Arise

The review identifies several reasons for moral dilemmas arising from the mismatch between ethical standards and moral values in AI design, deployment, and usage.

Moral dilemmas linked to perceptions of AI's agency: Users' perceptions of AI's moral agency can reduce personal accountability because of lower moral intensity on an ethical issue when AI systems perform tasks autonomously (Ebrahimi & Hassanein, 2021; Telkamp & Anderson, 2022; Zhang et al., 2021). According to Dahiyat (2021, p. 60), "intelligent software agents are capable of independent action rather than merely following instructions. [...]. This is why difficulties arise in deciding who should be responsible for the actions and mistakes such agents make.". The author argues that 'one-size-fits-all' regulation that does not recognize differences in AI's agency and users' corresponding perceptions can lead to a gap between legal standards (deontology) and responsibility practices (virtue ethics) (Dahiyat, 2021).

Moral dilemmas linked to different moral foundations: Ethical frameworks often emphasize principles like accountability and transparency, but Telkamp and Anderson (2022) argue that different stakeholders may prioritize different moral foundations, resulting in diverse judgments about the same AI-driven decision, wherefore AI systems often involve trade-offs between competing moral foundations (Telkamp & Anderson, 2022). Badea and Artus (2022, p. 3) argue that "[i]t is in principle impossible to specify any rule in such a way that it cannot be misinterpreted and what we mean by a rule can never be unambiguously represented." They build upon Wittgenstein's (1918) *Tractatus logico-philosophicus*² and describe this issue as the "interpretation problem," where abstract language fails to capture the complexity of real-world ethical concerns. These challenges demonstrate a gap between standard ethical frameworks and the diverse moral values of stakeholders.

Moral dilemmas linked to different moral maxims: Another potential source of moral dilemmas lies in striking a balance between strong regulation of AI, which may reduce users' motivation to adopt it, and weak regulation that fosters AI innovations that in the long term might maximize overall good but may comprise data privacy and risks misuse by malicious entities (Fernandes et al., 2020; Ratoff, 2021). Fernandes et al. (2020) argue that strong regulation could mandate AI technologies to adhere to utilitarian principles (utilitarianism), potentially conflicting with users' interests (virtue ethics). Conversely, a lack of regulation could spur innovation but also invade privacy issues, prompting non-AI adopters to demand stricter regulation (deontology). Thus, finding ethical standards that balance user preferences (related to their moral values) is a challenge.

Moral dilemmas linked to the absence of holistic guidance for different stakeholders that appear throughout AI's lifecycle: Another reason for moral dilemmas may stem from the absence of clear operational guidelines for the different roles of stakeholders being part of the AI lifecycle. For instance, machine learning (ML) developers focus on data preparation and model training, with efforts centered on developing de-biasing solutions (Gerdes, 2022). However, their exposure to AI ethics beyond the technology's design phase is limited. In comparison, Ibáñez and Olmeda

² Translated by David Pears and Brian McGuinness. London: Routledge, 1921.

(2022) found that managers rather focusing on norm-compliance and guidelines for employees, while governmental representatives focus on addressing unintended consequences for society. End-users may rather be aware of potential discrimination when using AI. Integrating ethics into AI technologies for all stakeholders throughout AI's lifecycle requires interdisciplinary approaches that consider relevant ethical standards before, during, and after implementation, including appropriate ethical measures (Gerdes, 2022).

4.3. Context-specific Moral Dilemmas: Understanding the Conflicts' Ontology

The review reveals that AI applications frequently generate moral dilemmas, often unnoticed in daily life, beyond well-known cases like the classical trolley problem of autonomous driving (Shank & Gott, 2020; Zhang et al., 2021). Table 1 provides an overview of context-specific moral dilemmas, showing how differing ontological views on AI and human agency shape perceptions and ethical conflicts..

Table 1. Overview of context-dependent potential moral dilemmas

Ontological and normative ethical assumptions	Potential moral dilemmas and conflict description	Top-down and bottom-up approaches for ethical human-AI interaction
<i>Recruitment and selection systems (e.g., decision-support tools, management advice)</i>		
<i>Material entity:</i> AI searches for candidates, screens and prioritizes applicants, automates first-round interviews, and recommends top candidates by predicting cultural fit (Fritts & Cabrera, 2021). This increases the agency of AI in recruitment but may dehumanize the process, treating applicants as a mere means to an end in a Kantian sense (Fritts & Cabrera, 2021). As Fritts and Cabrera (2021, p. 797) argue, using “artificial values” (substantial ontology) in hiring could be morally objectionable in itself. <i>Social entity:</i> The “machine heuristics” reflects employees’ perception that AI is more neutral than humans, even though they may still react to AI’s moral violations (Hong et al., 2020). Recruiters’ perceived distance from the applicants affected by their decisions can reduce moral intensity (relational ontology), potentially harming employee-employer relationships (Ebrahimi & Hassanein, 2021; Fritts & Cabrera, 2021).	<i>Should AI predictions be used to make more neutral application decisions even if it means that artificial values decide upon the life of candidates?</i> Here, the balance lies between neutrality (<i>deontology</i>, making objective decisions free from human bias) and preserving humanity (<i>virtue ethics</i>, empathy and contextual judgment by human recruiters). This conflict is rooted in substantial ontology, as it concerns the substitution of human values with artificial ones.	<i>Top-down approaches:</i> • Limit AI’s role in decisions that impact other people’s lives (Fritts & Cabrera, 2021) • Guide employees to consider all stakeholders (Fritts & Cabrera, 2021) • Increase moral recognition to decision subjects and reducing acceptance of discriminatory AI recommendations (Ebrahimi & Hassanein, 2021) <i>Bottom-up approaches:</i> • Develop AI systems aligned with societal values (Hong et al., 2020) • Address value conflicts on a case-by-case basis (Fritts & Cabrera, 2021)
<i>Conversational agents (e.g., Siri or Alexa)</i>		
<i>Material entity:</i> Despite AI’s programmed principles, its gamified interactions can alter discourse, sometimes in unintended ways, representing relational ontology (Fritts & Cabrera, 2021). For instance, Microsoft’s Tay.AI, a Twitter bot that quickly learned and replicated Nazi ideology, led to a disturbing interaction (Shank & Gott, 2020). Thus, AI agents “have the capability to make consequential decisions about communication on their own without necessarily needing input from a principal user and, in so doing, may profoundly shape how meaning is interactionally constructed” (Endacott & Leonardi, 2022, p. 464). Following Endacott and Leonardi (2022), AI agents can make autonomous decisions without direct user input (substantial ontology). <i>Social entity:</i> Respondents in the study of Shank and Gott (2020) described their smartphone’s AI apps as independent, attributing some degree of agency to the technologies (substantial ontology). When a certain app failed to meet expectations, respondents reclaimed their own agency to make more careful selections (substantial ontology).	<i>Should AI prioritize user convenience or ensure thorough consent to prevent inappropriate content exposure or the violation of personal rights?</i> Here, the balance lies between the greatest happiness principles (<i>utilitarianism</i>, maximizing user satisfaction) and the protection of personal rights and specific groups of people (<i>deontology</i>, protecting personal rights, especially for vulnerable groups). For instance, one participant in the study of Shank and Gott (2020, p. 1641) described the following situation: “Two children were in the room, ages 3 and 6. At one point, Alexa misunderstood one of our questions, and began reading an erotic story. I’m not sure what emotion the kids felt, but they were clearly disturbed and confused.” Here, prioritizing user rights protection is crucial, though, in other contexts, users may prefer optimized convenience, highlighting a nuanced challenge.	<i>Top-down approaches:</i> • Raise awareness of everyday situations in which moral dilemmas occur beyond the prominent examples (Shank & Gott, 2020) <i>Bottom-up approaches:</i> • Increase diplomacy with respect to problematic AI outputs; apologies to customers for misleading or unsatisfying AI outputs can help ensure those see the organization as competent (Endacott & Leonardi, 2022) • Increase knowledge of AI’s operating processes to aid understanding of what leads to certain outcomes and thus maintain agency based on knowledge and explanations (Endacott & Leonardi, 2022)

Crime detection algorithms (e.g., algorithms that build risk scores for prisoners; AI assistants that predict crime hotspots)		
<i>Material entity:</i> Following Martin (2019, p. 838), “algorithms are not neutral but value-laden in that they (1) create moral consequences, (2) reinforce or undercut ethical principles, or (3) enable or diminish stakeholder rights and dignity.” Crime detection algorithms act autonomously, and their implications affect individuals’ lives (Martin, 2019). Thus, their implications are related to humans (relational ontology). <i>Social entity:</i> Organizational representatives remain responsible, as they are the ones designing, deploying, and using AI technologies that are value-laden in nature (Martin, 2019). For instance, the company that is developing crime detection algorithms remains responsible to some extent, and the institutions that use such algorithms for public safety goals, such as the police, are also responsible (relational ontology).	<i>Should public institutions (e.g., policy) deploy an AI algorithm to predict crime if it risks reinforcing disproportionately targeting certain groups of people?</i> Here, the balance lies between maximizing public safety (<i>utilitarianism</i>, maximizing overall safety) and the risk of unfair treating certain groups (<i>virtue ethics</i>, disproportionately targeting certain demographic groups). This conflict pertains to relational ontology, as it addresses dilemmas that arise when AI outputs affect individuals’ lives and considers how institutions deploying AI should respond to these challenges.	<i>Top-down approaches:</i> • Implement regulatory oversight to ensure that ethical implications of algorithms meet societal standards (Martin, 2019) • Hold institutions accountable for the ethical implications of their algorithms, regardless of the technologies’ transparency or complexity (Martin, 2019) <i>Bottom-up approaches:</i> • Engage with the community affected by AI technologies to learn their concerns (Martin, 2019) • Allow stakeholders to understand AI’s processes and predictions by communicating transparently the underlying mechanisms that lead to certain predictions (Martin, 2019)
Medical AI assistants (e.g., brain tumor detection and segmentation tool, medical treatment adviser)		
<i>Material entity:</i> When AI assistants undertake diagnostic tasks, prescribe medicines, or make other medical treatment decisions, the performance of these tasks depends mainly on these systems (substantial ontology during the operational processes) (Lebovitz et al., 2021; Yokoi et al., 2021). AI technologies can work solely with what Lebovitz et al. (2021) call the “know-what aspects of knowledge” that limit the system’s outputs and usefulness, particularly in highly sensitive domains such as healthcare. <i>Social entity:</i> Healthcare is a knowledge-intensive and highly sensitive domain. The expertise of physicians and medical staff is developed over years as they accrue rich know-how knowledge through practice (substantial ontology during the operational processes) (Lebovitz et al., 2021). Thus, it is important to assess the outputs of medical AI assistants	<i>Should a medical AI assistant be able to prioritize treatment for patients with higher survival probabilities over those with severe conditions but lower survival probabilities?</i> Here, the balance lies between beneficence (<i>utilitarianism</i>, maximizing the lives saved and improving overall health outcomes through the allocation of medical resources where they are most likely to result in successful treatments) and equality expectations (<i>virtue ethics</i>, respecting the intrinsic value of each individual and avoid discrimination based on health status). This conflict can be grounded in substantial ontology because of the distinct and independent aspects of knowledge that differentiate medical AI assistants (know-what) from medical staff (know-how).	<i>Top-down approaches:</i> • Foster societal debate on AI’s benefits and risks in the context of medical treatment to enhance trust in the use of AI to support healthcare tasks and raise awareness of its limitations (Yokoi et al., 2021) <i>Bottom-up approaches:</i> • Do not rely on a single “ground truth” diagnosis, as the complexity and nuances of real-world medical practices require multifaceted know-how (Lebovitz et al., 2021) • Do not rely on AI’s outputs alone, as this impedes learning processes; instead, recognize the importance of human experts’ know-how in contexts where uncertainty and nuanced judgment are essential for making critical decisions (Lebovitz et al., 2021)

based on human experts' tacit knowledge when the underlying situation is uncertain (Lebovitz et al., 2021).

- Enhance understanding of the needs of patients, physicians and other medical staff, and the way the AI can and cannot address those needs (Yokoi et al., 2021)

Online shopping assistants (e.g., AI-based self-services in retail, online brand community detectors)

Material entity:

Human-like characteristics of AI shopping assistants increase customers' intention to engage in moral behavior (i.e., moral intention, such as reporting errors) (Giroux et al., 2022). This illustrates relational ontology rooted in anthropomorphic AI design and their underlying nudges (Wu et al., 2020).

Social entity:

Moral behavior and customer feelings of guilt decrease when they consider the AI assistant as less human-like (Giroux et al., 2022). When AI assistants aim to influence certain behaviors, punishment messages for misbehaving customers are effective in the short term but often reduce customers' future purchases (Wu et al., 2020). This illustrates how human behavior is linked to the nudges of online shopping assistants (relational ontology).

Should people be encouraged to view AI shopping assistants as more human-like to promote more sustainable shopping behavior, even if this could lead to misplaced guilt and responsibility?

Here, the balance lies between influencing moral behavior positively (*virtue ethics*, nudging customers to make more sustainable purchasing decisions) and transparency (*deontology*, emphasizing transparency about the capabilities and limitations of AI shopping assistants to avoid unrealistic expectations). This conflict can be grounded in relational ontology due to the impact of the anthropomorphic design of AI shopping assistants on human moral behavior.

Top-down approaches:

- Avoid “bad character animations” when designing human-like mental capacities (Giroux et al., 2022)
- Consider the impact of anthropomorphism of AI technologies on customer morality (Giroux et al., 2022)

Bottom-up approaches:

- Trigger moral values and foster moral identity among communities with AI technologies' mechanisms to enhance ethical behavior rather than implementing penalizing mechanisms (Wu et al., 2020)
- Use anthropomorphism to enhance customer moral actions, but be aware of the “uncanny valley” phenomenon, where increasingly the human-like characteristics of technologies can elicit feelings of discomfort when they reach a certain point of near-realism (Giroux et al., 2022)

Algorithmically managed platforms (e.g., algorithmic managed work, such as Uber)

Material entity:

Algorithmically managed platforms aim to influence user behavior through diverse nudging mechanisms based on collecting vast amounts of data from the user's environment (e.g., from several devices, their GPS navigation, usage behavior, and sensor data) (Liu et al., 2021). Thus, there is user dependency on the AI's outcomes within the platform's environment, which can be linked to relational ontology.

Social entity:

To succeed within algorithmically managed platforms, users must engage with an artificial value set (Fritts & Cabrera, 2021). Ratings, real-time monitoring, and incentives may

Should algorithmically managed platforms use nudges (e.g., surge pricing) and constant real-time monitoring to reduce users' moral hazards, despite the potential for opportunistic behavior?

Here, the balance lies between influencing moral behavior positively (*virtue ethics*, employing nudges and real-time monitoring transparently to encourage users to behave more ethically) and opportunistic behavior (*utilitarianism*, exploiting loopholes in algorithms to maximize personal gain). This conflict

Top-down approaches:

- Implement regulatory restrictions to avoid fraud (Liu et al., 2021)
- Fully inform users on the platform's mechanisms and provide concrete data policies as a platform provider (Liu et al., 2021)

Bottom-up approaches:

- Avoid biases that could unjustly penalize certain users. There is a risk of reinforcing existing inequalities or creating new ones through biased data and algorithmic decisions (Liu et al., 2021)

reduce the possibility of engaging in moral hazard (Liu et al., 2021). However, mechanisms such as surge pricing in ride-sharing apps can also incentivize drivers to behave opportunistically (Liu et al., 2021). This also can be linked to relational ontology.	is rooted in relational ontology as the algorithm’s mechanism is directly linked to certain human behavior.	Consider the potential for ML techniques to detect opportunistic behavior (Liu et al., 2021)

Autonomous vehicles (e.g., autonomous driving, autonomous aircraft)

Material entity:

Autonomous vehicles use pre-crash algorithms to make decisions about collision avoidance, enhancing safety, and reducing damage (Cunneen et al., 2019). This predetermined, calculated approach minimizes human error or biased decisions, even though the AI lacks moral reasoning (Cunneen et al., 2019). When faced with unavoidable accidents, autonomous systems require pre-set moral decision-making frameworks independent of the driver (substantial ontology) (Cunneen et al., 2019).

Social entity:

In unavoidable traffic incidents, human drivers have only a moment to make instinctual, sometimes morally complex decisions, leading to moral shortcuts (substantial ontology) (Cunneen et al., 2019). Rhim et al. (2020) introduce three moral-reasoning types of drivers: moral altruists who prioritize the safety of all parties (utilitarianisms), moral non-determinists who make decisions based on the specific context (virtue ethics), and moral deontologist who base their decisions on traffic rules (deontology). As automation reduces manual control, operators may become ethically disengaged, shifting accountability onto vehicle systems, creating “moral buffers” that distance humans from their actions (Holford, 2022).

Should responsibility for autonomous transportation be

distributed among all entities to increase safety despite the risk of ethical disengagement and potential failures in unprecedented emergencies?

Here, the balance lies between the maxim of minimizing overall harm (*utilitarianism*, prioritizing actions that minimize harm by distributing responsibility among all entities involved to collectively enhance safety and reliability of autonomous vehicles) and potential disagreement over values and spread responsibilities for unforeseen outcomes (*virtue ethics*, distributing responsibilities might lead to ethical disengagement, where entities may deflect responsibility). This dilemma can be linked to relational ontology since it relates responsibility for AI operations and outcomes to the diverse stakeholders involved.

Top-down approaches:

- Empower operators with control and information to influence autonomous systems as needed (Holford, 2022)
- Be aware that different countries prefer different normative structures (Ebina & Kinjo, 2021)
- Establish culture-specific guidelines for autonomous vehicle practitioners (Rhim et al., 2020)

Bottom-up approaches:

- Ensure active human engagement to maintain awareness of vehicle status and operations (Holford, 2022)
- Incorporate social welfare functions to address complex moral decisions, like the trolley problem, in unavoidable situations (Ebina & Kinjo, 2021)

4.4. Managing Ethical Human-AI Interaction

The review outlines three pathways for integrating top-down standards with bottom-up values:

Clarifying misperceptions of AI's (moral) agency: There is a need to address misleading human perceptions of AI's moral agency through training and standardized AI contracts. Despite AI technologies lacking moral agency in the way humans possess it, people often ascribe some level of morality to these machines (Ratoff, 2021). Zhang et al. (2021) emphasize the importance of aligning AI's role with human ethical expectations to prevent misleading perceptions and ensure responsible integration into society. Dahiyat (2021) further suggests standardized AI contracts to ensure that AI acts according to different people's will to increase user information on legal statuses and responsibilities when interacting with these systems. Ibáñez and Olmeda (2022) recommend organizations to consider the entire lifecycle of AI technologies when providing ethical training for several stakeholders. Also, Wang et al. (2020) propose robust data governance frameworks to increase compliance with existing regulations.

Incorporating virtue ethics into the AI design and implementation: Badea and Artus (2022) argue that virtue ethics should be embedded into AI technologies using game theory logic to interpret given standards through external moral values. Similar to games, rules as ethical standards can establish a normative space of possibilities in which players interpret rules differently and make various decisions to achieve their individual moral objectives, which creates relational moral values such as "being trusted" (virtue ethics), extending mere principles of trustworthiness principles (deontology). Gerdes (2022) suggests another approach involving user-centered and participatory design methods to integrate domain-specific values into AI design. This approach builds on Friedman et al.'s (2017) idea of value-sensitive design, which aims to involve interdisciplinary stakeholders at various levels (conceptual, empirical, technical) of AI design to translate moral values (virtue ethics) into technically realizable design requirements. La Fors et al. (2019) further support this by advocating for a holistic approach that considers the entire lifecycle of AI technologies and involves diverse stakeholders.

Introducing regulatory oversights and multi-objective accountability: Ratoff et al. (2021) and Telkamp & Anderson (2022) highlight the need for regulatory frameworks to ensure the ethical development and use of AI technologies. Both articles argue that there is no 'one-size-fits-all' approach, as the perceptions of AI ethics vary across different contexts. Given the value-laden nature of AI, it is crucial to integrate multiple moral perspectives to create systems perceived as ethical across various cultural settings and application scenarios (Telkamp & Anderson, 2022). For instance, Ebina and Kinjo (2021) demonstrate that different countries favor different normative ethical approaches. The authors integrate three different types of social welfare functions—Bentham's, Nash's, and Rawl's type—into an automatic control model for optimal driving behavior within the context of autonomous vehicles to determine inequality aversion. Their findings indicate that while Bentham's type, which aligns with utilitarianism by saving as many people as possible, is preferred in the United States, Canada, Australia, and some Western European countries, African and Asian countries favor settings where the autonomous vehicle adheres to given rules, reflecting a deontological approach. Multi-objective approaches that share common defining characteristics of different ethical frameworks (e.g., utilitarian and deontological ones) may help to align AI decision-making standards more closely with human values, enabling better management of multiple conflicting objectives (Vamplew et al., 2018). Such regulatory frameworks would identify moral thresholds so that AI developers and organizations are compelled to focus on creating AI that makes decisions that are not only logical but also morally understandable and do not deviate from human moral standards (Hong et al., 2020; Ratoff, 2021).

5. Discussion

This section provides theoretical and practical contributions.

5.1. Theoretical Contributions

This work structures the findings of the reviewed literature along the entities of an STS, following the conceptualization of Leonardi (2012), illustrating its appropriateness in explaining the interplay between ethical standards and moral values.

Beginning with the **positive influence of moral values** (arising from the *social subsystem*) **on ethical standards** (guiding the *sociomaterial entanglement*): Certain work practices within the *social subsystem*, such as specific data preparation rules developed by AI designers based on their moral values, can be perceived as ethical by others, resulting in their adoption as best practices and thus guiding *sociomaterial entanglement* and transforming into broader ethical standards (e.g., Gerdes, 2022). This carries the risk, though, of unintentionally generalizing best practice examples in undesired ways (Constantinescu et al., 2021). Furthermore, attempts to integrate moral values of various stakeholders from the *social subsystem* into the design of AI technologies through user-centered and participatory methods ensure that ethical standards are better aligned with societal values (e.g., Badea & Artus, 2022; Gerdes, 2022; La Fors et al., 2019). However, embedded moral values of organizational representatives can make it challenging to develop universally applicable normative structures, resulting in value-laden ethical principles (Badea & Artus, 2022; Martin, 2019; Ratoff, 2021). This becomes evident when the ethical standards derived from one specific *social subsystem*, with its unique normative structure, are used to guide the *sociomaterial entanglement* between humans and AI that are embedded within another normative structure.

Continuing with the **positive influence of ethical standards** (in AI technologies as *material entities*) **on moral values** (of human users as *social entities*): High ethical standards in AI technologies can prevent the dehumanization of individuals, ensuring— by recognizing and respecting human dignity and the worth of every person (Fritts & Cabrera, 2021)—that people are not treated merely as a means to an end, reflecting Kantian ethics (Kant, 1991). Ethical standards from the *social subsystem* can positively influence the moral values of users, as regulatory or organizational guidelines can enhance moral recognition among employees, fostering a sense of responsibility and ethical conduct (Ebrahimi & Hassanein, 2021). Neglect of ethical standards by AI technologies can, paradoxically, also foster moral values of users by motivating them to provide explanations and apologies to those affected by wrong AI decisions, thus reinforcing ethical reflection and moral accountability (Endacott & Leonardi, 2022). Failures to meet ethical standards in AI design can further exacerbate negative moral outcomes.

Conversely, there can also be a **negative influence of ethical standards** (from the *social subsystem*) **on moral values** (of human users as *social entities*): Ethical standards from the *social subsystem* can also adversely affect user moral values as ethical frameworks often fail to accommodate fundamental disagreements about what constitutes ethical behavior (Telkamp & Anderson, 2022). Given that ethical standards are built through consensus, this approach imposes limitations by setting a minimal threshold for acceptable ethical standards, potentially lower than the higher moral values of the individuals involved (Constantinescu et al., 2021; Meijer et al., 2023). In addition, AI regulations that are too strict and accompanied by harsh judgments from the *social subsystem* can prompt the development of coping mechanisms or a search for alternatives (Wu et al., 2020), such as loopholes, ultimately resulting in a lack of willingness to take responsibility (Shank & Gott, 2020). Thus, ethical standards may not capture the diverse and complex nature of moral values from different stakeholders, potentially clash with individuals' moral blame behavior about wrongdoing (McDonald & Pan, 2020). Thus, while top-down integration of ethical standards through the *social subsystem* may work well within one *sociomaterial entanglement*, it may not be adequate within another entanglement.

When examining the **negative influence of ethical standards** (in AI technologies as *material entities*) **on moral values** (of human users as *social entities*), it becomes apparent that too much focus on certain ethical standards can have unintended consequences. For instance, excessive transparency in the mechanisms behind AI algorithms may provoke opportunistic behavior among users if algorithms are linked to financial benefits (Liu et al., 2021). Moreover, AI technologies that adhere strictly to principles without considering contextual nuances may undermine the expertise and know-how of human users, which are crucial in uncertain situations (Fritts & Cabrera, 2021; Lebovitz et al., 2021). Therefore, while AI technologies designed with rigid principles may be

beneficial for some human users within one *sociomaterial entanglement*, that rigidity could lead to immoral behavior when the entanglement is within another context.

5.2. Practical Contributions

The findings offer three practical contributions.

Nuanced integration of both approaches: Developing hybrid approaches that effectively balance top-down and bottom-up measures is essential to bridge gaps between community-driven ethics and regulatory oversights. Measures should be complemented by educational programs and user training - otherwise top-down regulations risk alienating users (Endacott & Leonardi, 2022). Complex or opaque AI contracts may discourage engagement, leading users to ignore them entirely, thereby exacerbating misunderstandings about human-like AI systems.

Cultural and industrial contexts in AI applications matter: Balancing both approaches further requires careful consideration of the specific ethical demands inherent to different AI application contexts and cultural environments (Fritts & Cabrera, 2021). The industry of application and the prevailing cultural norms significantly influence ethical requirements, necessitating that AI practitioners address potential conflicts between different cultural factors when implementing AI technologies (Rhim et al., 2020). These conflicts might arise between regulatory restrictions and bottom-up approaches that alignment with the societal values of particular cultural communities (Fritts & Cabrera, 2021; Martin, 2019), wherefore recognizing ethical pluralism is crucial (Dahiyat, 2021). There are instances when localized ethical considerations are required, creating a tension between universal standards and community-driven ethics—such as when comparing the use of AI in healthcare decisions to its use in military applications.

Tailored steps for different stakeholders throughout the AI lifecycle: Ethical AI design, deployment, and utilization necessitate considering diverse stakeholders' perspectives on ethical AI practices at different stages of the AI lifecycle. *Developers* should integrate virtue ethics into AI models to align with core societal values. Participatory design methods and interdisciplinary stakeholder involvement are vital bottom-up strategies to translate moral values into technically realizable design requirements (Badea & Artus, 2022; Gerdes, 2022). *Managers* play a pivotal role in aligning AI deployment objectives with corporate social responsibility (CSR) practices (Wang et al., 2020). Given the immaturity of “responsible AI practices,” the establishment of AI ethics advisory boards as a top-down measure is recommended (Wang et al., 2020, p. 4966). *End-users* should actively engage with AI providers by voicing concerns and sharing experiences, informing bottom-up refinements to AI systems. *Governments* must implement top-down measures to educate users about their rights.

6. Conclusion

This study provides a comprehensive analysis of how ethical standards can either bolster or undermine moral values within organizations, and how these dynamics influence ethical AI design, deployment, and utilization. By examining the balance of top-down and bottom-up approaches, the review highlights the need for tailored, context-specific ethical frameworks.

References

- Aslan, A., Greve, M., & Lembcke, T.-B. (2022). Let's Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence. *Proceedings of the Americas Conference on Information Systems (AMCIS)*.
https://aisel.aisnet.org/amcis2022/sig_odis/sig_odis/10
- Badea, C., & Artus, G. (2022). Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents. In M. Bramer & F. Stahl (Eds.), *Artificial Intelligence XXXIX* (Vol. 13652, pp. 124–137). Springer International Publishing.
https://doi.org/10.1007/978-3-031-21441-7_9
- Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, 185(4), 725–740.
<https://doi.org/10.1007/s10551-023-05339-7>
- Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, 28(3), 801–831.
<https://doi.org/10.1086/345321>
- Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, 30(2), 195–218. <https://doi.org/10.1007/s11023-020-09525-8>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). MANAGING ARTIFICIAL INTELLIGENCE. *MIS Quarterly*, 45(3), 1433–1450.
<https://doi.org/10.25300/MISQ/2021/16274>
- Bilal, A., Wingreen, S., Sharma, R., & Jahanbin, P. (2021). Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics. *Proceedings of the Australasian Conference on Information Systems (ACIS)*.
- Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines*, 27(4), 591–608. <https://doi.org/10.1007/s11023-017-9448-z>

- Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26. <https://doi.org/10.1007/s10676-018-9448-6>
- Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. *Business Ethics Quarterly*, 33(1), 146–179. <https://doi.org/10.1017/beq.2021.42>
- Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, 23(3), 387–399. <https://doi.org/10.1007/s11019-020-09948-1>
- Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The Sociomateriality of Information Systems. *MIS Quarterly*, 38(3), 809–830.
- Champagne, M., & Tonkens, R. (2023). A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot harm. *Science and Engineering Ethics*, 29(4), 27. <https://doi.org/10.1007/s11948-023-00449-x>
- Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*, 23(4), 803–814. <https://doi.org/10.1007/s10676-021-09616-9>
- Cook, T. (2023). Robust Artificial Moral Agents and Metanormativity. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 162–169. <https://doi.org/10.1145/3600211.3604703>
- Córdova, P. R., & Vicari, R. M. (2022). Practical Ethical Issues for Artificial Intelligence in Education. In *Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal, August 31–September 2,*

- 2022, *Revised Selected Papers* (Vol. 1720, pp. 437–453). Springer Nature Switzerland.
<https://doi.org/10.1007/978-3-031-22918-3>
- Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2019). Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. *Applied Artificial Intelligence*, 33(3), 267–293.
<https://doi.org/10.1080/08839514.2018.1560124>
- Dahiyat, E. A. R. (2021). Law and software agents: Are they “Agents” by the way? *Artificial Intelligence and Law*, 29(1), 59–86. <https://doi.org/10.1007/s10506-020-09265-1>
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>
- Ebina, T., & Kinjo, K. (2021). Approaching the social dilemma of autonomous vehicles with a general social welfare function. *Engineering Applications of Artificial Intelligence*, 104, 104390. <https://doi.org/10.1016/j.engappai.2021.104390>
- Ebrahimi, S., & Hassanein, K. (2021). Decisional guidance for detecting discriminatory data analytics recommendations. *Information & Management*, 58(7), 103520. <https://doi.org/10.1016/j.im.2021.103520>
- Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1(1), 73–80. <https://doi.org/10.1007/s43681-020-00011-6>
- Endacott, C. G., & Leonardi, P. M. (2022). Artificial Intelligence and Impression Management: Consequences of Autonomous Conversational Agents Communicating on One’s Behalf. *Human Communication Research*, 48(3), 462–490. <https://doi.org/10.1093/hcr/hqac009>
- Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, 21(4), 403–418. <https://doi.org/10.1007/s10892-017-9252-2>

- Faulkner, P., & Runde, J. (2010). The Social, the Material, and the Ontology of Non-Material Technological Objects. *European Group for Organizational Studies (EGOS) Colloquium, Gothenburg*, 985, 4–8.
- Fernandes, P., Santos, F. C., & Lopes, M. (2020). Norms for Beneficial A.I.: A Computational Analysis of the Societal Value Alignment Problem. *AI Communications*, 33(3–6), 155–171. <https://doi.org/10.3233/AIC-201502>
- Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2), 63–125. <https://doi.org/10.1561/11000000015>
- Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, 23(4), 791–801. <https://doi.org/10.1007/s10676-021-09615-w>
- Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, 36(1), 2009222. <https://doi.org/10.1080/08839514.2021.2009222>
- Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>
- Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI. *Journal of Business Ethics*, 178(4), 1027–1041. <https://doi.org/10.1007/s10551-022-05056-7>
- Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, 26(1), 13. <https://doi.org/10.1007/s10676-024-09755-9>
- Greene, T., Copenhagen Business School, Shmueli, G., National Tsing Hua University, Ray, S., & National Tsing Hua University. (2023). Taking the Person Seriously: Ethically Aware IS Research in the Era of Reinforcement Learning-Based Personalization. *Journal of the*

- Association for Information Systems*, 24(6), 1527–1561.
<https://doi.org/10.17705/1jais.00800>
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Harzing, A.-W. (2005). *Journal Quality List*. [http:// www.harzing.com/](http://www.harzing.com/)
- Hawkins, W., & Mittelstadt, B. (2023). The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 261–270. <https://doi.org/10.1145/3593013.3593995>
- Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3), 101772. <https://doi.org/10.1016/j.jsis.2023.101772>
- Holford, W. D. (2022). An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control? *Journal of Business Ethics*, 176(1), 141–157. <https://doi.org/10.1007/s10551-020-04640-z>
- Hong, J.-W., Choi, S., & Williams, D. (2020). Sexist AI: An Experiment Integrating CASA and ELM. *International Journal of Human–Computer Interaction*, 36(20), 1928–1941. <https://doi.org/10.1080/10447318.2020.1801226>
- Hooker, J. N., & Kim, T. W. N. (2018). Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 130–136. <https://doi.org/10.1145/3278721.3278753>
- Horneber, D., & Laumer, S. (2023). Algorithmic Accountability. *Business & Information Systems Engineering*, 65(6), 723–730. <https://doi.org/10.1007/s12599-023-00817-8>
- Horváth, I. (2022). AI in interpreting: Ethical considerations. *Across Languages and Cultures*, 23, 1–13. <https://doi.org/10.1556/084.2022.00108>

- Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY*, 37(4), 1663–1687. <https://doi.org/10.1007/s00146-021-01267-0>
- Jones, C., Parker, M., & Bos, R. T. (2005). *Business Ethics. A Critical Approach*. Routledge Taylor & Francis Group.
- Kant, I. (1991). *The metaphysics of morals*. Cambridge University Press.
- Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2022, November 15). AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers. *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*. Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering. <http://arxiv.org/abs/2207.01493>
- Kieslich, K., Keller, B., & Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1), 205395172210929. <https://doi.org/10.1177/20539517221092956>
- La Fors, K., Custers, B., & Keymolen, E. (2019). Reassessing values for emerging big data technologies: Integrating design-based and application-based approaches. *Ethics and Information Technology*, 21(3), 209–226. <https://doi.org/10.1007/s10676-019-09503-4>
- Latour, B. (1992). Where Are The Missing Masses? The Sociology of a Few Mundane Artifacts. *Shaping Technology/Building Society: Studies in Sociotechnical Change*, 1, 225–258.
- Lebovitz, S., Levina, N., & Lifshitz-Assa, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What. *MIS Quarterly*, 45(3), 1501–1526. <https://doi.org/10.25300/MISQ/2021/16564>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>

- Leonardi, P. M. (2012). Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How Are They Different? Do We Need Them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and Organizing* (pp. 24–48). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199664054.003.0002>
- Liu, M., Brynjolfsson, E., & Dowlatabadi, J. (2021). Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis. *Management Science*, 67(8), 4665–4685. <https://doi.org/10.1287/mnsc.2020.3721>
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, 160(4), 835–850. <https://doi.org/10.1007/s10551-018-3921-3>
- Mayer, A.-S., Strich, F., University of Passau (Germany), Fiedler, M., & University of Passau (Germany). (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, 239–257. <https://doi.org/10.17705/2msqe.00036>
- Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt.
- McDonald, N., & Pan, S. (2020). Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–19. <https://doi.org/10.1145/3415218>
- Meijer, A., Wiarda, M., Doorn, N., & Van De Kaa, G. (2023). Towards responsible standardisation: Investigating the importance of responsible innovation for standards development. *Technology Analysis & Strategic Management*, 1–15. <https://doi.org/10.1080/09537325.2023.2225108>

- Millar, J. (2016). An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars. *Applied Artificial Intelligence*, 30(8), 787–809. <https://doi.org/10.1080/08839514.2016.1229919>
- Mingers, J., & Harzing, A.-W. (2007). Ranking journals in business and management: A statistical analysis of the Harzing data set. *European Journal of Information Systems*, 16, 303–316.
- Mirbabaie, M., Brendel, A. B., Faculty of Business and Economics, Technische Universität Dresden, Hofeditz, L., & Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1), 726–753. <https://doi.org/10.17705/1CAIS.05034>
- Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE INTELLIGENT SYSTEMS*.
- Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. *AI & SOCIETY*, 36(2), 429–443. <https://doi.org/10.1007/s00146-020-01002-1>
- Mutch, A. (2013). Sociomateriality—Taking the wrong turning? *Information and Organization*, 23(1), 28–40. <https://doi.org/10.1016/j.infoandorg.2013.02.001>
- Neubert, M. J., & Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. *Business Horizons*, 63(2), 195–204. <https://doi.org/10.1016/j.bushor.2019.11.001>
- Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-023-01635-y>
- Orlikowski, W. J. (2010). The sociomateriality of organisational life: Considering technology in management research. *Cambridge Journal of Economics*, 34(1), 125–141. <https://doi.org/10.1093/cje/bep058>
- Orlikowski, W. J., & Scott, S. V. (2008). 10 Sociomateriality: Challenging the Separation of Technology, Work and Organization. *Academy of Management Annals*, 2(1), 433–474. <https://doi.org/10.5465/19416520802211644>

- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1), 34–47. <https://doi.org/10.1109/TTS.2020.2974991>
- Pruski, M. (2023). Ethics framework for predictive clinical AI model updating. *Ethics and Information Technology*, 25(3), 48. <https://doi.org/10.1007/s10676-023-09721-x>
- Puri, A. (2020). Moral Imitation: Can Algorithm Really Be Ethical? *Rutgers Law Record*, 48(1), 47–57.
- Ratoff, W. (2021). Can the predictive processing model of the mind ameliorate the value-alignment problem? *Ethics and Information Technology*, 23(4), 739–750. <https://doi.org/10.1007/s10676-021-09611-0>
- Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: An empirical examination of the moral individual. *Journal of Applied Psychology*, 92(6), 1610–1624. <https://doi.org/10.1037/0021-9010.92.6.1610>
- Rhim, J., Lee, G., & Lee, J.-H. (2020). Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of Korea and Canada. *Computers in Human Behavior*, 102, 39–56. <https://doi.org/10.1016/j.chb.2019.08.010>
- Roberts, J. S., & Montoya, L. N. (2022). *Contextualizing Artificially Intelligent Morality: A Meta-Ethnography of Top-Down, Bottom-Up, and Hybrid Models for Theoretical and Applied Ethics in Artificial Intelligence* (arXiv:2204.07612). arXiv. <http://arxiv.org/abs/2204.07612>
- Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, 33(1), 100925. <https://doi.org/10.1016/j.hrmr.2022.100925>

- Saariluoma, P., & Leikas, J. (2020). Designing Ethical AI in the Shadow of Hume's Guillotine. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), *Intelligent Human Systems Integration 2020* (Vol. 1131, pp. 594–599). Springer International Publishing. https://doi.org/10.1007/978-3-030-39512-4_92
- Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2), 103174. <https://doi.org/10.1016/j.im.2019.103174>
- Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7(1), 119. <https://doi.org/10.1057/s41599-020-00614-8>
- Shank, D. B., & Gott, A. (2020). Exposed by AIs! People Personally Witness Artificial Intelligence Exposing Personal Information and Exposing People to Undesirable Content. *International Journal of Human-Computer Interaction*, 36(17), 1636–1645. <https://doi.org/10.1080/10447318.2020.1768674>
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, 109–124. <https://doi.org/10.17705/1thci.00131>
- Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, 31(2), 74–87. <https://doi.org/10.4018/JDM.2020040105>
- Spiekermann, S. (2021). What to Expect From IEEE 7000: The First Standard for Building Ethical Systems. *IEEE Technology and Society Magazine*, 40(3), 99–100. <https://doi.org/10.1109/MTS.2021.3104386>
- Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, 45, 105661. <https://doi.org/10.1016/j.clsr.2022.105661>

- Swanepoel, D. (2021). The possibility of deliberate norm-adherence in AI. *Ethics and Information Technology*, 23(2), 157–163. <https://doi.org/10.1007/s10676-020-09535-1>
- Telkamp, J. B., & Anderson, M. H. (2022). The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. *Journal of Business Ethics*, 178(4), 961–976. <https://doi.org/10.1007/s10551-022-05057-6>
- Templier, M., & Paré, G. (2018). Transparency in literature reviews: An assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, 27(5), 503–550. <https://doi.org/10.1080/0960085X.2017.1398880>
- Tigard, D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics*, 30(3), 435–447. <https://doi.org/10.1017/S0963180120000985>
- Unver, M. B. (2023). Rebuilding ‘ethics’ to Govern AI: How to re-set the boundaries for the legal sector? *International Conference on Artificial Intelligence and Law (ICAIL)*. <https://doi.org/10.1145/3594536.3595156>
- Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1), 27–40. <https://doi.org/10.1007/s10676-017-9440-6>
- VHB. (2015). *VHB-JOURQUAL3 Ranking*. Verband der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft e.V. (VHB). <https://vhbonline.org/service/vhb-jourqual/vhb-jourqual-3/gesamtlste>
- Wang, Y., Xiong, M., & Olya, H. (2020). *Toward an Understanding of Responsible Artificial Intelligence Practices*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2020.610>
- Wickson, F., & Forsberg, E.-M. (2015). Standardising Responsibility? The Significance of Interstitial Spaces. *Science and Engineering Ethics*, 21(5), 1159–1180. <https://doi.org/10.1007/s11948-014-9602-4>

- Wilson, R. A., & Land, M. K. (2021). Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review*, 52(3), 1029–1194.
- Winfield, A. (2019). Ethical Standards in Robotics and AI. *Nature Electronics*, 2(2), 46–48.
- Wittgenstein, L. (1918). *Tractatus logico-philosophicus*.
- Wu, W., Huang, T., & Gong, K. (2020). Ethical Principles and Governance Technology Development of AI in China. *Engineering*, 6(3), 302–309.
<https://doi.org/10.1016/j.eng.2019.12.015>
- Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human–Computer Interaction*, 37(10), 981–990.
<https://doi.org/10.1080/10447318.2020.1861763>
- Yu, B., Vahidov, R., & Kersten, G. E. (2021). Acceptance of technological agency: Beyond the perception of utilitarian value. *Information & Management*, 58(7), 103503.
<https://doi.org/10.1016/j.im.2021.103503>
- Zarnoufi, R., Boutbi, M., & Abik, M. (2020). AI to Prevent Cyber-Violence: Harmful Behaviour Detection in Social Media. *International Journal of High Performance Systems Architecture*, 9(4), 182–191.
- Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *Journal of Artificial Intelligence Research*, 71.
<https://doi.org/10.1613/jair.1.12895>

Appendix I.

Table I.1 gives an overview of the format screening.

Table I.2. Format Screening

Exclusion criteria	Exclusion description	Eliminated articles
Non-scientific and peer-reviewed articles (n=437)	Gray literature, book chapters, book reviews, dissertations, tutorials, bachelor's or master's workshops, posters, letter of editors, special issues, symposiums, talks, magazine, and articles of scientific conferences. Articles not being peer reviewed.	ACM Digital Library: 123 articles; AIS Electronic Library: 147 articles; Business Source Complete: 36 articles; Science Direct: 25 articles; Web of Science: 106 articles
Incomplete articles (n=60)	Short papers or research-in-progress papers, e abstracts (articles with fewer than ten pages) without full-text availability.	ACM Digital Library: 2 articles; AIS Electronic Library: 17 articles; Business Source Complete: 7 articles; Science Direct: 3 articles; Web of Science: 31 articles
Non-English articles (n=7)	Articles not completely written in English are excluded.	ACM Digital Library: 1 article; AIS Electronic Library: 0 articles; Business Source Complete: 4 articles; Science Direct: 1 article; Web of Science: 1 article
Stand-alone literature reviews (n=149)	Standalone literature review papers following systematic approach are excluded.	ACM Digital Library: 8 articles; AIS Electronic Library: 19 articles; Business Source Complete: 34 articles; Science Direct: 25 articles; Web of Science: 63 articles
Total: 653 articles		

Appendix II.

Even though not being part of the VHB JOURQUAL-3 Ranking, articles that have been published in the academic journals or conferences that either refer to “Artificial Intelligence” or “Ethics” in a technological context are included. These are:

- Artificial Intelligence and Law
- AI Communication
- AI Practitioners
- Computers and Education: Artificial Intelligence
- Engineering Application of Artificial Intelligence
- Journal of Experimental & Theoretical Artificial Intelligence#
- Artificial Intelligence in Medicine
- Ethics and Information Technology
- Science and Engineering Ethics
- AI & Society

Appendix III.

Table III.1 gives an overview of the relevance screening.

Table III.3. Relevance Screening

Exclusion criteria	Exclusion description	Eliminated articles
No considerations standards and values	Articles that did not contribute to the discussion of the interplay between ethical standards and moral values	ACM Digital Library: 13 articles; AIS Electronic Library: 1 article;

(n=177)	(e.g., by analyzing moral dilemmas)—by either Business Source Complete: 96 articles; focusing solely on ethical principles (e.g., AI Science Direct: 34 articles; regulation) or solely on moral perceptions (e.g., Web of Science: 34 articles virtues)—were excluded. According to the moral values represent personal beliefs about “right” or “wrong,” shaped by an individual's societal, or philosophical influences, and root virtues like justice, courage, and prudence. These values guide personal decision-making and are compared, ethical standards are formalized or shared norms developed through consensus stakeholders (e.g., developers, managers, policymakers) to ensure responsible behavior, especially in contexts like AI. These standards provide guidance on what is ethically acceptable, often aligning with deontological ethics, which focus on duty, norms, and compliance with regulations	
No AI technologies (n=164)	Articles that did not mention AI technologies following the definition provided at the beginning of the Introduction as technologies that have evolved from simple rule-based systems to complex systems capable of informing, self-learning, predicting, and guiding decisions and actions in various contexts, both professional and personal.	ACM Digital Library: 16 articles; AIS Electronic Library: 1 article; Business Source Complete: 70 articles; Science Direct: 17 articles; Web of Science: 60 articles
No human-AI interaction focus (n=61)	Articles that did not consider the interaction between humans and AI following the idea of the social entanglement of Leonardi (2012)—by either solely on the technological artifact or solely on the management of human workers—were excluded	ACM Digital Library: 3 articles; AIS Electronic Library: 0 articles; Business Source Complete: 52 articles; Science Direct: 2 articles; Web of Science: 4 articles
Total: 402 articles		

Appendix IV.

Table IV.1 gives an overview of the final sample.

Table IV.4. Final Sample

Authors (Year)	Title	Outlet
Badea & Artus (2020)	Morality, Machines, and the Interpretation Problem: A Value-Aligned Wittgensteinian Approach to Building Moral Agents	Artificial Intelligence
Cunneen et al. (2020)	Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma	Applied Artificial Intelligence
Dahiyat (2019)	Law and Software Agents: Are They “Agents” by the Way	Artificial Intelligence and Law
Ebina & Kinjo (2020)	Approaching the Social Dilemma of Autonomous Vehicles: A General Social Welfare Function	Engineering Applications of Artificial Intelligence
Ebrahimi & Hassani (2021)	Decisional Guidance for Detecting Discriminatory Data Analytics Recommendations	Information & Management Science
Endacott & Leonard (2020)	Artificial Intelligence and Impression Management: Considerations for Human Communication Researchers	Human Communication Research
Fernandes et al. (2020)	Norms for Beneficial AI: A Computational Analysis of the Value Alignment Problem	AI Communication
Fritts & Cabrera (2020)	AI Recruitment Algorithms and the Dehumanization Problem	Ethics and Information Technology
Gerdes (2022)	A Participatory Data-centric Approach to AI Ethics by Design	Applied Artificial Intelligence

Giroux et al. (2022)	Artificial Intelligence and Declined Guilt: Retailing Morali	Journal of Business Ethics
	Comparison Between Human and AI	
Holford (2022)	An Ethical Inquiry of the Effect of Cockpit Automation on	Journal of Business Ethics
	Responsibilities of Airline Pilots: Dissonance or Meaningfi	
	Control?	
Hong et al. (2020)	Sexist AI: An Experiment Integrating CASA and ELM	International Journal of Hu
		Computer Interaction
Ibáñez & Olmeda (	Operationalising AI Ethics: How Are Companies Bridging	AI & Society
	between Practice and Principles? An Exploratory Study	
Lebovitz et al. (202	Is AI Ground Truth Really True? The Dangers of Training	MIS Quarterly
	Evaluating AI Tools Based on Experts' Know-What	
La Fors et al. (2022	Reassessing Values for Emerging Big Data Technologies:	Ethics and Information Tec
	Integrating Design-Based and Application-Based Approac	
Liu et al. (2021)	Do Digital Platforms Reduce Moral Hazard? The Case of U	Management Science
	Taxis	
Martin (2019)	Ethical Implications and Accountability of Algorithms	Journal of Business Ethics
Ratoff (2021)	Can the Predictive Processing Model of the Mind Amelioræ	Ethics and Information Tec
	Value-alignment Problem?	
Rhim et al. (2020)	Human Moral Reasoning Types in Autonomous Vehicle M	Computers in Human Beha
	Dilemma: A Cross-cultural Comparison of Korea and Canæ	
Shank & Gott (202	Exposed by AIs! People Personally Witness Artificial Intel	International Journal of Hu
	Exposing Personal Information and Exposing People to Un	Computer Interaction
	Content	
Telkamp & Anders	The Implications of Diverse Human Moral Foundations for	Journal of Business Ethics
(2022)	Assessing the Ethicality of Artificial Intelligence	
Vamplew et al. (20	Human-aligned Artificial Intelligence is a Multiobjective P	Ethics and Information Tec
Wang et al. (2019)	Toward an Understanding of Responsible Artificial Intellig	Hawaii International Confe
	Practices	System Sciences
Wu et al. (2021)	FairPlay: Detecting and Deterring Online Customer Misbel	Information System Resear
Yokoi et al. (2021)	Artificial Intelligence Is Trusted Less than a Doctor in Mec	International Journal of Hu
	Treatment Decisions: Influence of Perceived Care and Valæ	Computer Interaction
	Similarity	
Zhang et al. (2022)	Artificial Intelligence and Moral Dilemmas: Perception of	Journal of Experimental So
	Decision-Making in AI	Psychology
Total: 26 articles		

Appendix V.

Table V.1 provides a snapshot of the coding scheme.

Table V.5. Snapshot of Coding

Code	Coding Rule	Example
Relational ontology (<i>de code</i>)	Relational ontology posits that entities do not have properties independent of their relations with other entities. In the context of human-AI interaction, relational ontology suggests that the nature and behavior of both humans and AI systems are shaped by their dynamics and that roles are co-constructed through the sociomaterial entanglement.	“In fact, intelligent software agents are not merely following instructions. They further exhibit high levels of moral intelligence, and autonomy accounts for which their actions are not always anticipated, intended, or known by users.” (Dahiyat, 2021, p. 60)

Substantial ontology (<i>de</i> <i>code</i>)	Substantial ontology posits that entities have an independent and intrinsic existence that is characterized more on the decisions made by stable properties. In the context of human-AI interaction, it does not mean that the decision-making process is based on the substantial ontology would treat AI technologies as accepted unconditionally.” (Hofmann et al., 2020, p. 7)
Top-down approach (<i>in</i> <i>code</i>)	Top-down approaches refer to the top-down establishment of collective ethical standards and norms accounting for the fact that an advisory board can be established at several layers of AI governance guidelines, policies, and laws. The board can provide advice and guidance to the Board of Directors. (Wang et al., 2020, p. 4968)
Bottom-up approach (<i>in</i> <i>code</i>)	Bottom-up approaches refer to the bottom-up integration of societal values accounting for diverse moral communities. A participatory and centric approach to AI Ethics by engaging domain experts during the system design process and, furthermore, help raise organizational awareness of the challenges related to data-driven knowledge generation.” (Gerde et al., 2020, p. 770)

Paper VIII

Hammerschmidt, T.; Stolz, K. & Posegga, O. (2025). *Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI*. Behaviour and Information Technology, (BIT), <https://doi.org/10.1080/0144929X.2025.2500451>.

Bridging the Gap: Inequalities that Divide Those Who Can and Cannot Create Sustainable Outcomes with AI

Completed Research Paper – Accepted Version

Abstract

The widespread use of AI technologies impacts individuals, organizations, and societies in ways that warrant more comprehensive investigation. Inequalities in benefiting from AI advancements may lead to an “AI divide”—a gap between those who can generate successful and sustainable outcomes with AI and those who cannot. Although human-AI interactions can lead to various outcomes, such as increased efficiency and improved task performance, this study emphasizes the importance of fostering sustainable outcomes that account for the social, ethical, and environmental impacts of AI usage. In doing so, this work demonstrates how achieving success in AI utilization can also contribute to achieving the United Nations’ Sustainable Development Goals (SDGs). To better understand the disparities that divide individuals, organizations, and societies that can generate successful and sustainable outcomes with AI from those who cannot, we analyze and explain how several social, technical, and sociotechnical inequalities limit sustainable outcomes of human-AI interaction by integrating insights from a systematic literature review on AI inequalities with existing research on the digital divide and sociotechnical system theory. This approach conceptualizes the AI divide as a multi-level and circular construct, linking performance-related outcomes with sustainable development goals to advance individual, economic, environmental, and societal well-being through human-centered AI.

Keywords: Digital Divide, Digital Inequalities, Sustainable Development Goals, AI Adoption, Human-Centered Artificial Intelligence

1. Introduction

Artificial Intelligence (AI) describes technologies with the capability “to perform cognitive functions that we associate with human minds” (Rai et al., 2019, p. iii). While these technologies are increasingly being integrated into various aspects of human life, their implications still need to be better understood (Dwivedi et al., 2023; Ozmen Garibay et al., 2023).

It is important to understand that AI is not a singular concept; rather, it is used as an umbrella term for various techniques and technologies, including—among others—machine learning, natural language processing, computer vision, and robotics, each with distinct applications and implications (Benbya et al., 2020). Understanding these different types of AI is essential when examining the positive and negative outcomes of using them. A differentiated perspective is particularly vital in understanding how the use and adoption of such technologies relate to broader digital inequalities (Lutz, 2024). When investigating these inequalities, the United Nations (2020) Sustainable Development Goals (SDGs) can serve as a guiding framework for systematically approach research on inequalities related to AI, such as on gender equality (SDG#5).

Over the past decades, research on inequalities induced by specific technologies has been heavily influenced by research on the digital divide, which refers to inequalities between individuals, organizations, and societies in accessing and effectively using digital technologies (Broadbent & Papadopoulos, 2013; Wei et al., 2011). As AI continues to evolve, it gives rise to a more complex form of inequality, referred to as the AI divide (Carter et al., 2020). This AI divide results from disparities in the ability to develop, access, and benefit from AI technologies, further exacerbating existing digital inequalities (Carter et al., 2020; Lutz, 2024). To fully understand the AI divide, it is crucial to examine how different types of AI contribute to disparities across various domains.

It is essential to investigate why the benefits of using AI, such as enhanced efficiency, are not equitably distributed among individuals, organizations, and societies (Goralski & Tan, 2020). For instance, from an individual perspective, tech-savvy individuals may be more qualified to lead AI adoption (Kolbjørnsrud et al., 2017) or, from a societal one, legal restrictions might make some countries more able to leverage AI benefits compared to others (Wu, 2022). From an economic perspective, some companies possess greater financial resources to invest in AI than others (Mikalef & Gupta, 2021). These inequalities divide those who can and cannot create sustainable outcomes with AI, thereby obstructing progress toward several Sustainable Development Goals of the United Nations (2020). Notably, these include SDG#10, which aims to reduce inequalities within and among countries to ensure equal opportunities; SDG#4, which focuses on enhancing equality of opportunities through education; SDG#5, aimed at ending gender inequality and empowering especially girls and women; and SDG#16, which concerns inclusiveness (Goralski & Tan, 2020; Wamba et al., 2024). Deepening

our understanding of how and why inequalities result in the unequal distribution of AI benefits is, therefore, imperative (Lutz, 2024).

The traditional conceptualization of the digital divide does not adequately capture the sustainability implications of AI's increased autonomy, which may lead to novel and unexpected consequences for individuals, organizations, and societies (Gupta et al., 2023; Kitsara, 2022; Wamba et al., 2024). While AI supports certain SDGs, such as affordability and accessibility, it may also inhibit progress toward other targets, such as equality and well-being, leading to inequalities (Ozmen Garibay et al., 2023). To devise more equitable and inclusive ideas on how to overcome these inequalities, digital divide concepts must be adapted to current AI developments and aligned with sustainability research (Ozmen Garibay et al., 2023; Vassilakopoulou & Hustad, 2023). Clearly distinguishing between different types of AI and their roles in exacerbating or mitigating digital inequalities is a crucial step in addressing these challenges.

We advocate for understanding the AI divide as a function of inequalities that create disparities between individuals, organizations, and societies in terms of their ability to access and utilize AI to achieve multiple outcomes simultaneously. These outcomes encompass not only performance-related results, such as increased efficiency (e.g., Lin et al., 2021; Venkatesh, 2022), but also sustainable outcomes, including the development of inclusive AI solutions that consider social, ethical, and environmental factors (e.g., Kar et al., 2022; Leslie, 2019; Shneiderman, 2020). Success in AI implementation can also support the achievement of the United Nations' SDGs, and vice versa. By integrating performance-related outcomes with sustainability objectives, we prioritize human needs and values, consistent with the idea of human-centered AI (HCAI), which promotes the well-being of individuals, communities, and societies (Andersen et al., 2022; van Berkel et al., 2022).

This work further acknowledges that outcomes of access to and usage of AI technologies can sometimes be contradictory. For instance, the widespread accessibility of user-friendly generative AI technologies such as ChatGPT can reduce barriers to AI use, thus narrowing the AI divide in terms of greater usage efficiency and greater inclusiveness (Choudrie et al., 2022; Dwivedi et al., 2023; Seeber et al., 2020). This accessibility, though, can lead to non-mindful usage and thus to negative consequences, such as decreased well-being resulting from reduced human control or overreliance on AI's outputs (Carolus et al., 2023; M. Li & Suh, 2022; Mirbabaie et al., 2022; Seeber et al., 2020). In light of this, we must increase our understanding of how inequalities arise within AI contexts at different levels of access and usage, and concerning different outcomes, how those inequalities are interrelated, and how to overcome them to contribute to sustainable human-AI interaction outcomes according to the United Nations' SDGs.

Our research question (RQ) is: *How can we understand inequalities that limit sustainable outcomes of human-AI interaction and thus create an AI divide between individuals, organizations, and societies?*

To answer our research question, we conduct a systematic theoretical literature review to conceptualize the AI divide from a sustainability development perspective, drawing on sociotechnical systems theory (SST; Leonardi, 2012). Our goal is to contribute to research by enhancing knowledge about inequalities that arise with increased adoption of AI, as called for by Rajagopal and Vedamanickam (2019) and Carter et al. (2020). We enhance the theoretical foundations of research on the AI divide through a conceptualization that follows SST and involves sustainability, following the call of Vassilakopoulou and Hustad (2023). Based on our literature review findings, we offer recommendations for practitioners on using AI for social good to support the SDGs and spell out some avenues for future work on the AI divide.

2. Theoretical Background

This section illustrates why AI transformation necessitates contextualizing the digital divide concept (Section 2.1) and why SST is employed to structure the theoretical review findings in a conceptualization of the AI divide (Section 2.2).

2.1 The Need for Contextualizing Digital Divide Research

The digital divide is prominent in Information Systems (IS) research on inequality in the context of human-technology interaction (Robinson et al., 2015). This research spans multiple levels. Riggins and Dewan (2005) investigated access and usage, the first two levels of the digital divide. Following the literature on the digital divide, the access divide, or first-level divide, differentiates between individuals with and without access to digital technologies; the usage divide, or second-level divide, pertains to the skills that distinguish individuals who can and cannot use digital technologies (Ragnedda, 2017; van Deursen & Helsper, 2015), Wei et al. (2011) extended the digital divide framework of Riggins and Dewan (2005) empirically by adding the outcome divide as a third level; it encompasses inequalities in creating beneficial outcomes through human-technology interaction. With the widespread standardization of Internet access and the increasing usability of digital technologies, the concept of the digital divide has increasingly expanded from access to usage and from usage to outcomes, leading to a focus on creating tangible outcomes (Blank & Lutz, 2018; Scheerder et al., 2017). Figure 1 illustrates the digital divide framework of Wei et al. (2011) assuming a linear relationship between the three divide levels. It also outlines that the digital divide can be examined from an individual, organizational, or global perspective.

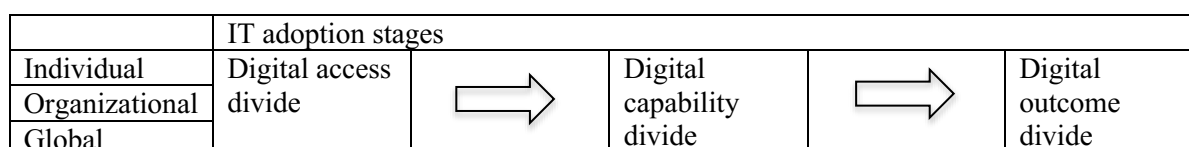


Figure 1. Three-Level Digital Divide Framework by Wei et al. (2011)

Various perspectives on the interrelation of these three digital divide levels exist within the digital divide literature. Ragnedda (2017) posits that the third level involves the advantages derived from the

first two levels, suggesting a linear progression from access to usage to outcomes. Van Deursen and Helsper (2015) highlight the necessity of shifting focus from merely considering usage skills differentiating groups that effectively can use technologies or cannot to the differences in the benefits and outcomes of using digital technologies. Blank and Lutz (2018), for instance, found that elderly users benefit more from the Web than younger users due to a higher potential for overcoming inequalities. This suggests that the outcome divide of elderly users might be reduced to a greater extent than for younger users, as the Web may help overcome previous limitations elderly users faced with respect to participating in social and daily life (e.g., due to immobility). Elderly users can utilize the Web for telehealth services, social networking, and accessing government services, significantly enhancing their quality of life. Greater participation in various aspects of life narrows the overall outcome divide for seniors.

Merely reducing the digital divide in terms of usage (usage divide), however, does not necessarily lead to positive outcomes of human-AI interaction. For instance, increased AI usage can sometimes lead to stress and reduced well-being, thereby diminishing the overall benefits of human-AI interaction (Blank & Lutz, 2018). This illustrates that positive usage (and, thus, a smaller usage divide) can sometimes lead to contradictory outcomes (such as a greater outcome divide). These contradictory outcomes are particularly important when it comes to AI technologies, given that AI's greater autonomy and complexity can lead to several intended or unintended consequences for societies, individuals, and organizations (Baidoo-Anu & Owusu Ansah, 2023; Dwivedi et al., 2023). This necessitates differentiating between inequalities that hinder the creation of various outcomes as a result of AI utilization, such as between inequalities that hinder the creation of successful outcomes through AI utilization (in terms of increased performance and efficiency) and those that hinder sustainable outcomes (in terms of understanding AI's implications and creating inclusive AI outcomes). This highlights the need to contextualize digital divide concepts (e.g., Keskin & Vermeulen, 2024; Kitsara, 2022).

The need for a specific conceptualization of the AI divide can be illustrated through the four phases of digital transformation to AI transformation, following Davenport (2018), which leads to changes in inequalities as described in the digital divide literature, as seen in Appendix I. For instance, Klumpp (2017) calls for more research that examines the so-called algorithmic divide as a narrow concept of the AI divide, focusing on disparities caused by algorithms. Luttrell et al. (2020) emphasize the importance of researching the AI divide in light of AI's transformative impact on various domains. Carter et al. (2020) highlight the need to investigate AI's implications for the digital divide by establishing comprehensive AI divide research frameworks that synthesize potential dividing factors and visible and invisible implications of AI. Carter et al. (2020) and Lutz (2024) applied the concept of the digital divide as illustrated in Figure 1 with the linear three-level structure proposed by Wei et al. (2011) to establish a sociotechnical framework for the AI divide. Following their work, the AI divides

include the access divide (inequalities in access to AI in both private and non-private settings), usage divide (differences in skills, use, and participation in AI technology), and outcome divide (inequalities in effectively utilizing AI outcomes).

Recent research on the AI divide, much like studies on the digital divide, has been approached from various perspectives and refer to a global divide, social divide, and democratic divide, for instance (Norris, 2001). These include a societal or geographical perspective that examines how different nations or regions benefit from or are excluded from AI advances (e.g., through different digital connectivity standards); an organizational perspective focused on how institutions can leverage AI technologies; and an individual perspective that explores how individuals interact with and benefit from AI. For instance, Kitsara (2022) investigated how AI innovations can exacerbate socioeconomic gaps, highlighting the divide between industrialized and developing countries as well as between privileged and less-privileged organizations and individuals (e.g., those with limited financial resources or governmental support). In a similar vein, the World Economic Forum adopted a societal perspective, distinguishing between leading AI countries and emerging ones (Bughin & van Zeebroeck, 2018). Wu (2022) deployed a societal perspective by investigating the AI divide of China's society.

While much research has focused on one of these perspectives in isolation, it is essential to recognize their interdependencies. For instance, organizational decisions regarding AI implementation and training can significantly affect individuals' use of AI technologies, including for personal activities such as leisure (Chowdhury et al., 2023). This illustrates an interdependency between the individual and organizational perspectives. Factors such as education and cultural norms also impact how individuals perceive AI technologies and develop realistic narratives (Diederich et al., 2022). Hence, to investigate the AI divide of individuals, organizations, and societies, it is crucial to consider all possible interdependencies between those perspectives. A comprehensive analysis requires attention to how these different levels—societal, organizational, and individual—interact and influence one another.

2.2 The Need for SST to Contextualize AI Inequalities

Building on Vassilakopoulou and Hustad (2023), who illustrate the digital divide as a sociotechnical phenomenon and highlight the necessity of understanding how technologies intersect with social phenomenon to link digital divide research with sustainability, we propose that SST offers a robust framework for examining the inequalities that hinder sustainable outcomes in human-AI interactions. A sociotechnical system consists of an arrangement of interdependent social and technical entities (e.g., humans, technologies) that collectively shape outcomes within a given context, such as an organization or society (Leonardi, 2012). By emphasizing the interplay between these entities, SST provides a nuanced perspective on the complex interactions that influence AI-driven inequalities and their broader implications for sustainability.

SST is grounded in the broader theoretical lens of sociomateriality, which underscores the inseparability of technologies as material entities from the social practices of human actors as social entities (Orlikowski, 2007). Together, both entities form the sociomaterial entanglement (Orlikowski, 2007), which either can foster or hinder inequalities within human-AI interactions. Traditionally, technologies were viewed as passive tools used by humans; recent literature highlights how the increasing autonomy of AI technologies and their ability to actively shape interactions transform how technologies and humans shape each other and how they are shaped by human decision-making, organizational structures, and broader societal conditions (Endacott & Leonardi, 2022). Unlike conventional digital tools, AI systems do not merely mediate interactions but participate in them, influencing decision-making processes, organizational structures, and broader societal conditions (Endacott & Leonardi, 2022). This evolving relationship between humans and AI (referred to as sociomaterial practices) calls for a deeper exploration of how sociotechnical dynamics reinforce or mitigate inequalities.

In the context of AI, sustainable outcomes can be linked to ethical principles for human-AI interactions that generate long-term, ethically grounded benefits (Heyder et al., 2023). The sustainability of AI-driven interactions is not dependent solely on technological advancements, but is also shaped by the broader social subsystem in which AI is embedded. According to Leonardi (2012), the social subsystem encompasses norms, guidelines, institutional structures, cultural frameworks, and values—all of which play a critical role in shaping human-AI interactions. Within this subsystem, both non-human elements (e.g., organizational policies, and regulatory frameworks) and human-related elements (e.g., values, and cultural norms) create a nested structure that enforces sociotechnical dynamics. (Leonardi, 2012; Winter et al., 2014). These dynamics influence AI technologies as material entities and humans as social entities, operating at multiple levels, including societal, industrial, and organizational perspectives (Heyder et al., 2023). As a result, these interdependencies determine whether AI systems contribute to bridging or exacerbating digital inequalities, ultimately shaping what is commonly referred to as the AI divide (Carter et al., 2020).

AI-supported hiring processes offer an example of such sociotechnical dynamics in AI usage. Kelan (2024, p. 2) observed that human bias may be reproduced in AI-supported hiring, potentially leading to exclusion. Although AI hiring algorithms may enhance recruiters' performance—reducing the AI outcome divide for recruiters—they might simultaneously hinder inclusiveness for applicants as other social entities interact with the same AI technology, leading to a greater AI outcome divide among applicants. Recognizing different sociomaterial practices within AI usage cases can help us better understand the disparities in outcomes from human-AI interaction.

To summarize, SST serves as an appropriate theoretical framework for contextualizing the digital divide in human-AI interactions, enabling a deeper understanding of how sustainable outcomes emerge—or fail to emerge—within the interdependent arrangement of AI technologies and human

users. By structuring the findings of our systematic literature review within the SST framework, we aim to derive a comprehensive conceptualization of the AI divide. This approach aligns with Avgerou (2019), who advocates for a sociomaterial perspective in research on human-AI interactions, emphasizing the need to consider both technological and social influences when analyzing AI-related inequalities.

Figure 2 is an overview of how we apply Leonardi’s (2012) SST conceptualization to structure the systematic literature review findings within a conceptualization of the AI divide. Whereas Figure 1 assumes a linear progression of the AI divide across three levels—the first-level divide (access divide), second-level divide (usage divide), and third-level divide (outcome divide)—Figure 2 takes a different approach by examining the interplay between these divide levels, allowing for a more nuanced analysis of their interdependencies as revealed through the systematic literature review.

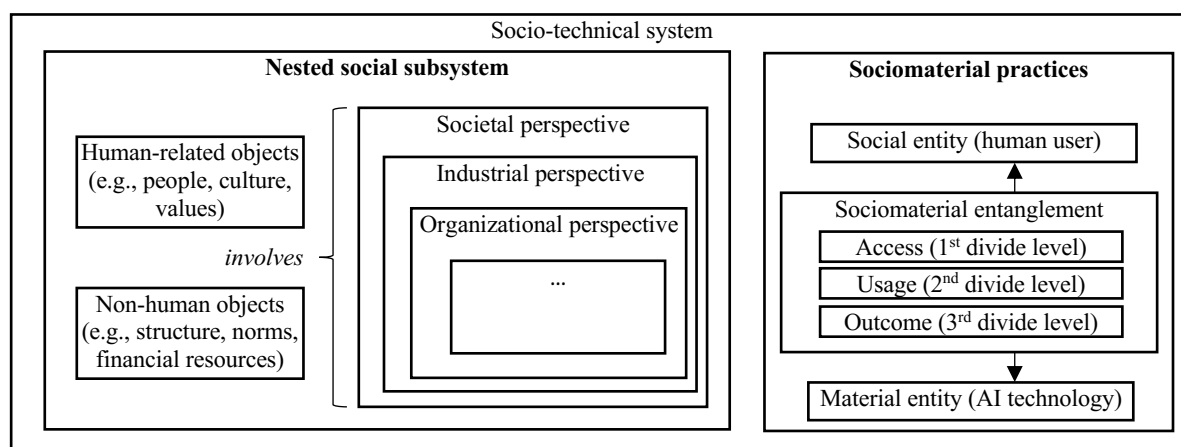


Figure 2: Adapted sociotechnical system of Leonardi (2012)

3. Research method

To derive a conceptualization of the AI divide and answer our RQ, we conduct a theoretical literature review to identify, describe, and transform relevant objects into a novel structure (Paré et al., 2015). We followed the steps of Templier and Paré (2018) for a systematic and transparent search process.

Literature search: We searched for relevant literature by querying the title, abstracts, and keywords—if available—for the keywords in the two search strings. The first string is: (“Artificial Intelligence Divide” OR “Artificial Divide” OR “AI Divide” OR “Algorithmic Divide”); the second string is: (“Artificial Intelligence” OR “generative AI” OR “AI-based digital assistant” OR “AI-based assistant” OR “AI assistant” OR “AI-based conversational agent” OR “AI agent” OR “AI technology”) AND (“inequalities” OR “inequality” OR “divide” OR “gap” OR “barrier”). We searched in the following databases: ACM Digital Library (ACM), AIS Electronic Library (AIS), Business Source Complete (BSC), IEEE Computer Society Digital Library (CSDL), Google Scholar (Scholar), Science Direct (SciDir), and Web of Science (WoS). The literature search, conducted in

July 2023 and updated in July 2024, yielded 8,709 overall hits. After eliminating duplicates (n=156), we yielded 8,553 unique hits.

Literature screening: Due to ongoing debate and the great evolution of the academic literature on AI technologies, we did not restrict our search to specific quality criteria (such as outlet rankings) or to certain time limits. We applied format criteria through format screening, eliminating non-scientific (e.g., non-peer-reviewed), non-English, and incomplete articles. If available, we used database filter options to eliminate articles that did not fulfill format criteria. We excluded standalone literature reviews. To summarize, 2,764 articles were eliminated due to format criteria, and 5,789 articles remained for relevance screening.

We applied further relevance criteria by eliminating articles that do not focus on an assistive purpose of AI technologies that do not make dedicated contributions to the conceptualization of the AI divide because they do not consider factors that differentiate between individuals, organizations, and societies that are or are not AI-capable, and that focus solely on the barriers to the adoption of AI for specific tasks without considering human-AI interaction within a sociotechnical system. In total, 5,743 articles were excluded, and 46 articles remained. Appendix II provides further details on format and relevance screening.

Backward and forward searches: We also conducted four rounds of backward and forward searches. For backward search, we checked the sources cited in each of the remaining articles of our final sample. For forward search, we looked for articles that refer to each of the remaining articles, using Google Scholar's "cited by" function. This identified 86 articles, which we further checked against the aforementioned format and relevance criteria. In all, 73 articles were eliminated for having been already screened (52 in the first round of backward and forward search, 18 in the second round, 3 in the third round, and 1 in the fourth round). The process was repeated until no additional relevant articles were identified. This added another 13 articles, resulting in a total of 59 articles as our final sample. Appendix III provides overviews of the final sample and its database distribution.

Figure 3 is an overview of the search process, following the steps of Templier and Paré (2018).

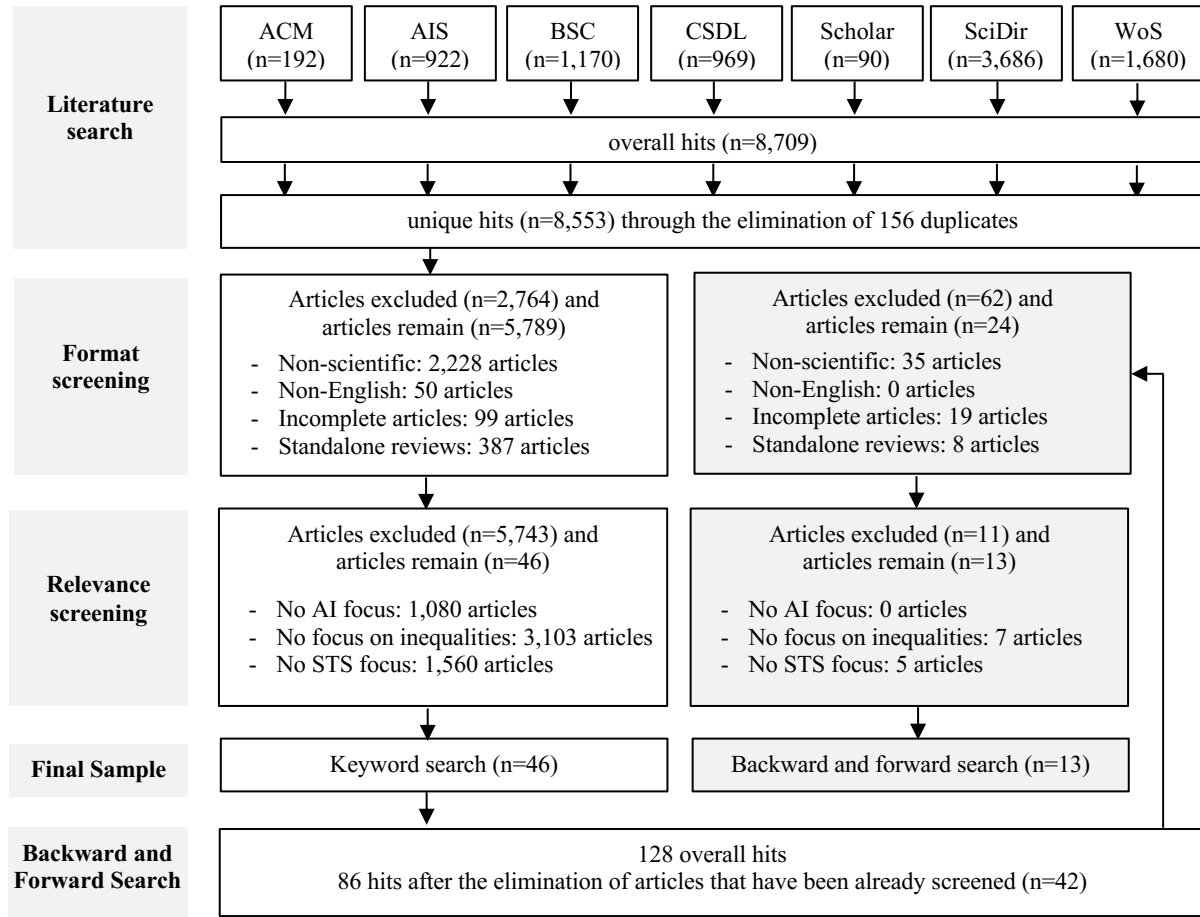


Figure 3: Search Process.

Literature analysis: Overall, the literature review yielded a final sample of 59 articles, which were downloaded as PDFs and analyzed through manual coding using Microsoft Excel. First, we structured the samples' meta information (e.g., regarding the research domain, method, ranking of the publication outlet) to provide a descriptive overview of the different article types within the final sample. We then opted for qualitative analysis to structure the content of relevant articles. We used the theoretical background to structure the findings deductively (Mayring & Fenzl, 2014), applying inductive coding when novel concepts were identified. We conducted parallel coding to guarantee consistency during the coding process and used discussion rounds among the coders. Appendix IV is a snapshot of coding.

To increase the quality of coding, we randomly chose a small subsample of the articles (n=58) for parallel training purposes of the coders before applying the exclusion criteria. The coding scheme was revised after a subsequent discussion. We used Cohen's Kappa, following Cohen (1960), to measure interrater agreement and increase our review reliability, obtaining a score of $\alpha=0.75$, which indicates a substantial agreement between coders (McHugh, 2012).

Conceptualizing: As illustrated in the Introduction, SST is used to structure our literature review findings. We use SST as a deductive foundation that is extended through the theorizing part of our

review, adopting an integrating type of conceptualization by following the guidelines for conceptual work proposed by MacInnis (2011). Integrating involves the synthesis of the main findings into the structure of the SST conceptualization from Figure 2 (see Section 2.2). In doing so, the conceptualization should serve as an “architectural plan“ aggregating the discourse on AI inequalities to explain possible interdependencies through SST that can guide further empirical studies (MacInnis, 2011, p. 138).

4. Results

Our literature review offers novel insights into the drivers of inequalities limiting sustainable human-AI interactions in response to our research question: “*How can we understand inequalities that limit sustainable outcomes of human-AI interaction, creating an AI divide between individuals, organizations, and societies?*” We have incorporated these findings into the framework of SST and digital divide literature, as outlined in the theoretical background (Section 2), leading to the conceptualization of the AI divide presented in Figure 4 and then followed by a presentation of four main findings on the conceptualization of the AI divide. These are:

- Drivers of inequalities within **sociomaterial practices** (section 4.1): why compensation alone will not suffice [right side of Figure 4: referring to factors of the material and social entity that can drive inequalities]
- The AI divide as a circular concept (section 4.2): a multi-level approach to analyzing AI inequalities within the **sociomaterial entanglement** [right side of Figure 4: referring to the different divide levels]
- The significance of the outcome divides within the **sociomaterial entanglement** (section 4.3): addressing contradictions in AI outcome inequalities [right side of Figure 4: referring to the outcome divide as the third divide level]
- The role of nested **social subsystems** (section 4.4): AI inequalities arise from several perspectives [left side of Figure 4]

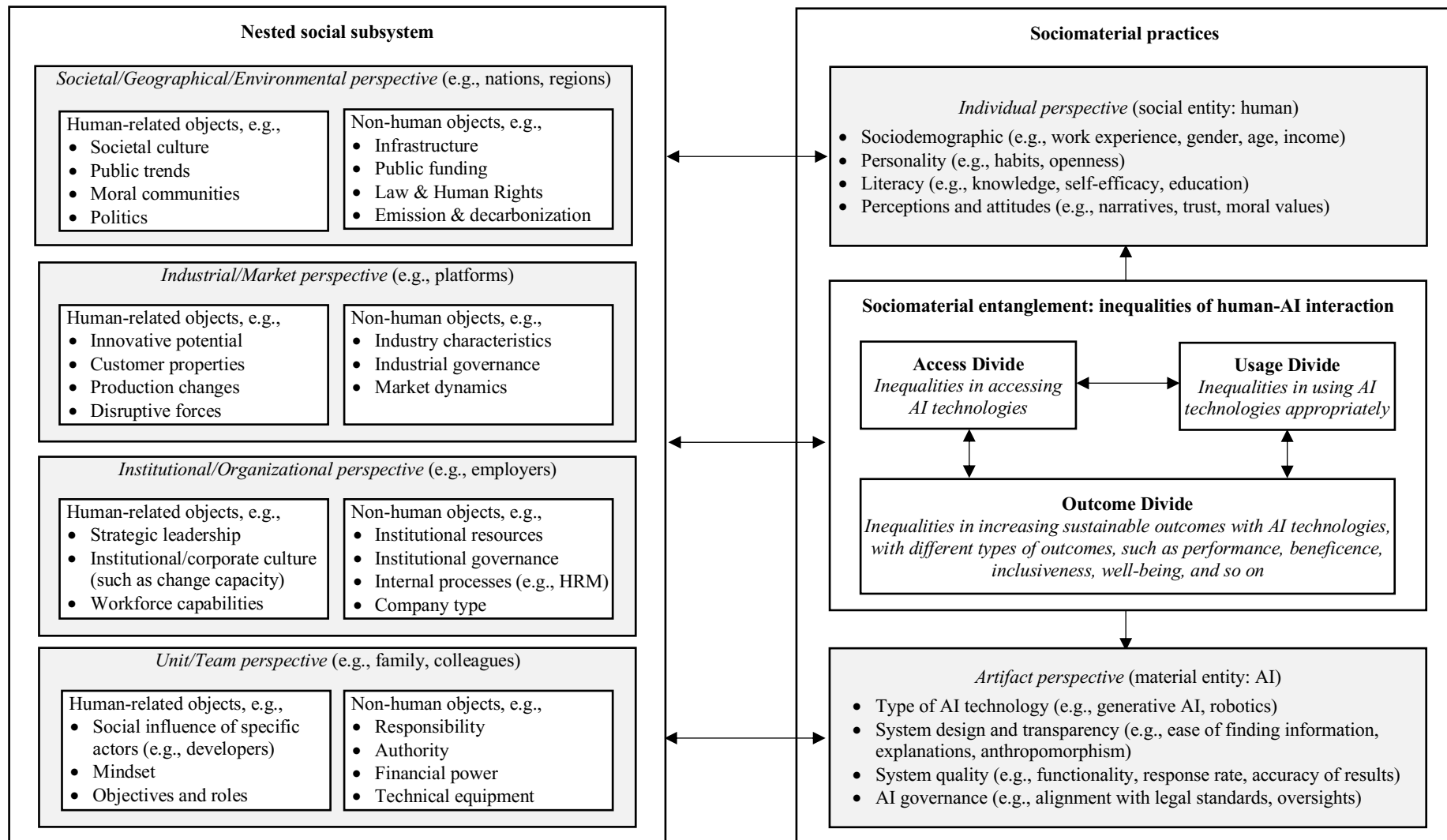


Figure 4: Conceptualization of the AI divide

4.1 Drivers of inequalities within sociomaterial practices: why compensation alone will not suffice

The review identifies several drivers of inequalities within the sociomaterial practices between AI technologies as material entities and human users as social entities (see the right side of Figure 4). Even if an individual human user experiences no inequalities personally, the presence of significant inequalities in the AI's design—such as a lack of transparency or inadequate ethical frameworks—can result in outcomes that are not sustainable (Nascimento & Meirelles, 2022). Efforts to compensate for human user inequalities, such as insufficient skills in handling AI technologies, by improving the quality of AI outputs and enhancing user experience (García de Blanes et al., 2022; Goldenthal et al., 2021; Kutz et al., 2022; Zhang et al., 2023) may offer some benefit. However, this approach may only partially succeed, perhaps helping overcome barriers to effective human-AI interaction but perhaps exacerbating the sustainable outcome divide by obscuring fundamental functionalities.

When examining **drivers of inequalities from the human user as the social entity**, our review primarily highlights sociodemographic inequalities, personality-related inequalities, skill and literacy inequalities, and perception-related inequalities.

Sociodemographic inequalities often involve gender disparities and differences in socioeconomic status and education. For instance, Aydin et al. (2023) found that gender inequalities arise because women's circumstances are often overlooked when automating tasks with AI, such as their overrepresentation in part-time or care-related roles, which influences their career progression. This is underpinned by Young et al. (2023), who suggest that patterns in AI and data science echo the gender gaps present in the broader workforce. Their study reveals that female professionals using AI are more likely to work in traditionally feminized industries such as healthcare and are underrepresented in technological and manufacturing sectors. In addition, female participants reported lower confidence in their AI abilities. However, the extent to which AI reinforces or mitigates these patterns remains underexplored, signaling a need for further research rather than assuming a predetermined effect. Young et al. (2023) argue that this situation leads to ethical issues related to social justice and economic opportunity, as leadership and technical roles driving AI design and deployment remain male-dominated.

W. Li et al. (2024), through a survey of 965 adults in the United States, found that certain sociodemographic groups, such as those who attended higher education institutions, are more likely to have access to and proficiency with AI technologies. The study also found that “males and younger generations were more knowledgeable about AI than females and older generations” (W. Li et al., 2024, p. 17). This is supported by Daepf and Counts (2024, p. 1), who argued that “education is the strongest positive predictor of rates of search for generative AI tooling.” Thus, highly educated

people tend to seek out novel AI technologies, which may further widen the social AI divide between educated and less-educated individuals, as those with more education and technological awareness continue to advance their AI proficiency at a faster rate. Although tech-savvy individuals may be better equipped to lead AI adoption, there may still be challenges to creating sustainable outcomes. For example, younger employees in the generation that grew up with advanced technologies and can be considered tech-savvy may produce outputs quickly using AI, but with their focus on efficiency may pay less attention to the consequences of AI use (Xing et al., 2021). In contrast, leaders responsible for AI implementation strategies must consider the broader implications of AI, including its long-term effects, but also must decide when, where, and how AI should be implemented; incorrect decisions may impede AI's potential to create added value to an organization (Alsheibani et al., 2020). Therefore, sociodemographic inequalities can lead to varied outcomes in human-AI interactions.

Personality-related inequalities revealed in the review were related to usage behavior (e.g., habits), the commonality of AI within professional and private environments, and certain personality traits of the user to which the technology may adapt and incorporate as properties (e.g., Diel et al., 2021; Gignac & Szodorai, 2024). In addition, Mou and Xu (2017) investigate the relation between personality traits (openness, agreeableness, conscientiousness, extraversion, and neuroticism) and the communication attributes that can be observed in human-AI and human-human interaction. Overall, users communicate differently within the human-human than in the human-AI context, such as being more open, extroverted, conscientious, agreeable, and self-disclosing toward AI compared to humans (Mou & Xu, 2017).

Skill and literacy inequalities are closely related to the digital divide literature, where the usage divide is often framed as a skill divide or capability divide (e.g., Carter et al., 2020). As far back as 2013, Tseng and Ting (2013, p. 463) emphasized the crucial role of AI in the context of knowledge economies in several fields, such as “problem reasoning and solving, machine learning, network structures, and knowledge processing systems.” Most authors in the final sample draw on the definition by Long and Magerko (2020) when defining AI literacy, incorporating skills and knowledge as integral components of the concept (e.g., Celik, 2023; Oeldorf-Hirsch & Neubaum, 2023; Q. Wang et al., 2024). For example, Celik (2023) builds on Long and Magerko's (2020) definition by describing AI literacy as “a new set of abilities and competencies that enable people to utilize AI efficiently and ethically in daily life” (p. 102026). Our literature review underscores the importance of AI capabilities in reducing barriers to AI implementation (Bérubé et al., 2021; Goldenthal et al., 2021). Q. Wang et al. (2024), in their survey of 1,088 people in The Netherlands, examined the literacy and attitudes of vulnerable groups in AI-influenced online news and entertainment environments. They found that “the most vulnerable groups with the lowest levels of AI knowledge and skills were predominantly older, with lower levels of education and privacy

protection skills” (p. 15). This finding highlights the interdependencies between various social drivers that can exacerbate human inequalities and the fact that knowledge management is highly important for organizations, institutions, and society (Grundspenkis, 2007).

Similarly, Oeldorf-Hirsch and Neubaum (2023) identified such interdependencies, noting that inequalities often reinforce rather than compensate for one another. For instance, they argued that while some individuals may possess high self-efficacy and cognitive skills for effectively engaging with social media algorithms, their literacy often falls short when it comes to understanding the broader implications. The authors specifically point out that certain groups, such as LGBTQ+ communities, “face additional challenges in terms of their literacy, which is necessary to navigate algorithms in constantly evolving social systems” (Oeldorf-Hirsch & Neubaum, 2023, p. 28). This observation suggests that AI literacy must go beyond merely using algorithms efficiently to encompass an awareness of the social and ethical implications of their use. Oeldorf-Hirsch and Neubaum (2023) advocate for enhancing user literacy by raising awareness of potential biases and improving developer literacy to ensure the responsible design of social media algorithms and thus prevent the exacerbation of social inequalities. This shift reflects a broader move from focusing solely on efficiency to integrating sustainable outcomes into the concept of literacy, illustrating that AI literacy can significantly influence human-AI interaction outcomes (e.g., C.-H. Wang & Wu, 2022; Yang et al., 2021).

Perception-related inequalities involve how individuals perceive AI’s intelligence in comparison to human intelligence, which can shape certain narratives (e.g., Gignac & Szodorai, 2024) and affect the confidence necessary for engaging with specific behaviors (Kamoonpuri & Sengar, 2023; Xing et al., 2021). Strengthening people’s self-efficacy is crucial, as it encourages them to interact with novel technologies (Frick et al., 2021). Nascimento and Meirelles (2022) describe self-efficacy as a mindset that encompasses tacit beliefs and perceptions, stating that it involves “the efficacy expectation—the belief in one’s own ability to perform the behavior required to achieve outcomes—and the outcome expectancy—the individual’s evaluation that a particular behavior will produce a specific outcome” (p. 2). Since self-efficacy is closely linked to the performance-related aims of individuals, it plays a key role in narrowing the AI outcome divide. In addition, perceptions of AI’s intelligence capabilities relative to human intelligence are significant, as they influence how individuals define AI and, consequently, how they respond to it (Gignac & Szodorai, 2024). How people perceive AI also affects whether they develop realistic narratives about its impact on their lives (Diederich et al., 2022). Therefore, addressing perception-related inequalities is essential for fostering more informed and effective human-AI interactions, ultimately reducing the divide in AI outcomes.

When examining the **drivers of inequalities from AI technology as the material entity**, our review primarily highlights disparities related to different types of AI technologies, system design, quality, and underlying governance mechanisms.

Various types of AI technologies possess distinct characteristics that impact the AI divide in different ways (e.g., McElheran et al., 2023). For example, some technologies, like voice-based assistants, enhance usability and thus reduce the usage divide. Others, such as open-source generative AI tools, improve accessibility, narrowing the access divide. Additionally, certain technologies, like robots that augment human capabilities, can improve performance, reducing the outcome divide. Furthermore, technologies like automated-guided vehicles can increase inclusiveness for marginalized groups, thereby addressing the sustainable outcome divide (e.g., Gao, 2024; Goldenthal et al., 2021; McElheran et al., 2023; Uba et al., 2023). However, with these unique characteristics come different drivers that can exacerbate disparities.

For instance, Mannuru et al. (2023) noted that the benefits of generative AI, such as ChatGPT, are not equally accessible, particularly in developing countries where limited access to advanced technologies and inadequate infrastructure pose significant challenges. Similarly, Oomen et al. (2023) found that different types of AI technologies can evoke varying levels of aversion and appreciation among users. They argued that aversion can hinder the effective implementation of AI, thus limiting outcomes. However, this aversion often stems not from the technology itself but from the processes surrounding its use, such as issues of transparency and lack of user influence, which affect sustainable outcomes. For example, while Apple's Siri, a virtual assistant that processes voice commands, enhances human-AI interaction (thus reducing the usage divide through increased appreciation), it also raises privacy concerns when users do not expect their sensitive conversations to be recorded, leading to aversion and limiting sustainable outcomes (Oomen et al., 2023). Similarly, Clearview AI, a facial recognition tool used in law enforcement, scrapes images from social media and other online platforms, impacting the privacy of internet users worldwide and raising concerns about false-positive identifications (Oomen et al., 2023).

In contrast, AI technologies that do not involve personal or sensitive information, such as IBM Watson's participation in the game Jeopardy or Google DeepMind's AlphaGo playing the game Go, tend to garner higher levels of appreciation, even though the human-like performance of these AI systems can initially provoke aversion (Oomen et al., 2023). This aversion is often mitigated if the insights gained in these entertainment contexts can be applied to more practical and beneficial areas, such as medical research and practice, thereby reducing the AI outcome divide (Oomen et al., 2023). In summary, the interplay between various types of AI technologies and their unique technical characteristics underscores the complexity of addressing AI-related inequalities, where both the benefits and potential drawbacks must be carefully managed to reduce disparities across different divide levels.

The design of AI systems, particularly in terms of specific functionalities, transparency, and human likeness, plays a significant role in mitigating barriers associated with AI technologies. Transparency is especially vital in “high-risk [or] high-consequence” contexts in which AI systems are used for decision support (Kamoonpuri & Sengar, 2023, p. 4). Transparency and explainability foster a sense of control among users, which is crucial for encouraging AI adoption (Schaefer et al., 2021). Conversely, opaque AI data processing can create obstacles (Kutz et al., 2022; Pandl et al., 2021) that may exacerbate the AI divide. A lack of transparency can also hinder individuals’ ability to ensure the accuracy and reliability of AI outcomes (Xing et al., 2021), leading to accountability issues (Bérubé et al., 2021) and difficulties in assessing the correctness of AI outputs (Uren & Edwards, 2023), thereby compromising sustainable outcomes. Furthermore, anthropomorphic design can either enhance or hinder human-AI interaction. While some argue that a greater social presence through human-like design can reduce barriers to interaction (e.g., Al-Fraihat et al., 2023; Diel et al., 2021; Park & Humphry, 2019), others suggest that it may instead create challenges to achieving successful and sustainable outcomes with AI (e.g., Kamoonpuri & Sengar, 2023; Oomen et al., 2023). Thus, although “the affordability, multi-modality, and multilingual capabilities of these tools could also make them more accessible to diverse users” (Daepf & Counts, 2024, p. 1), and thus might reduce their access divide, how AI technologies are designed specifically impacts how people can create sustainable outcomes.

The quality of AI systems, including aspects such as underlying functionality, response rate, and result accuracy, is critical in achieving sustainable outcomes in human-AI interactions. Low-quality AI systems that are not user-friendly can impede these interactions, thereby widening the AI outcome divide (Nascimento & Meirelles, 2022). Such deficiencies in quality may arise from technical factors, including insufficient data availability (Jöhnk et al., 2021), data management challenges (Aydin et al., 2023; Hradecky et al., 2022), inadequate performance (García de Blanes et al., 2022; Zhang et al., 2023), and systems that are difficult to use, uncomfortable, or lack practical applicability (Goldenthal et al., 2021; Kutz et al., 2022). Additional factors contributing to low quality include lengthy response times (Kutz et al., 2022) and the failure of AI systems to integrate effectively with existing IT infrastructures (Kutz et al., 2022). Therefore, the quality of AI systems is crucial in bridging or widening the AI outcome divide, emphasizing the importance of addressing these technical shortcomings to ensure effective and sustainable human-AI interactions.

The underlying AI and data governance mechanisms are significant drivers that can impact the AI divide, with the existence of robust AI governance crucial for achieving sustainable outcomes. Mirbabaie et al. (2022, p. 89) note that effective AI governance involves “introducing explainable AI” and ensuring that trustworthy, responsible, and ethical people oversee AI implementation and application. To bridge the AI divide sustainably, it is essential to “establish domain responsibility for data [and] infrastructure” (Uba et al., 2023, p. 6136) alongside clear human accountability

(Mirbabaie et al., 2022). The absence of well-defined roles, principles, and responsibilities in AI governance can lead to privacy concerns and security issues (Al Shamsi et al., 2022; Diel et al., 2021; García de Blanes et al., 2022). Moreover, “AI-based systems are at risk for biased learning and unethical outcomes” (Jöhnk et al., 2021, p. 11), which can result in liability issues, challenges from unions or employees, and the need for legal regulation (Weinert et al., 2022). However, in the short term, AI governance may impede performance, since adhering to existing legal and ethical standards can slow down certain forms of innovation (e.g., Gao, 2024). However, in the long term, it may enhance overall performance by reducing the need for corrective measures to address past mistakes, such as unintended consequences or unforeseen events. This suggests that sustainable outcomes could, at least in part, offset the initial reduction in outcomes over time. Robust AI and data governance mechanisms are thus essential to mitigating risks and ensuring the ethical, secure, and sustainable implementation of AI technologies, thereby reducing the AI divide and, specifically, barriers to creating sustainable outcomes.

In conclusion, addressing the AI divide requires more than compensating for inequalities of human users as social entities; it also demands careful consideration of the drivers that shape AI technologies as material entities. Even when efforts are made to improve user experience and overcome personal inequalities, the underlying design and governance of AI systems can perpetuate or even exacerbate the divide, especially in terms of sustainable outcomes. As Park and Humphry (2019, p. 935) aptly noted, “digital assistance services that incorporate features of automation and/or algorithmic processing may add another layer of exclusion for those who are already socially and digitally excluded.” Therefore, overcoming AI-related inequalities requires a holistic approach that addresses both drivers of inequalities of the AI as a material entity and the human user as a social entity to ensure that AI benefits all users equitably as solely focusing on one entity will not compensate existing inequalities.

4.2 The AI divide as a circular concept: a multi-level approach to analyzing AI inequalities within the sociomaterial entanglement

To further elaborate on the circular relationships within the different divide levels of the sociomaterial entanglement (see the right side of Figure 4), our review highlights that inappropriate use of AI technologies, or not using them at all, can result in poor data quality or insufficient data volume. These create barriers to effective AI usage (usage divide) and can also limit access to adequate data (access divide) (Bérubé et al., 2021). Moreover, there may be daily life situations in which individuals do not recognize the presence of AI embedded within technology, as AI is often “automatically integrated into popular technologies” (Goldenthal et al., 2021, p. 2). This integration typically occurs through software updates that require no additional purchase or installation, making the use of AI less visible, particularly in examples within business-to-customers (B2C) applications like social media algorithms or generative AI tools. This contrasts with the more cost-intensive

implementation of business-to-business (B2B) AI applications within organizational processes (Kamoonpuri & Sengar, 2023). For instance, when users engage with social media to communicate with friends or participate in online activities, they may be unaware that the features they rely on and the recommendations they receive are driven by AI (Celik, 2023). Consequently, while these individuals may interact with the platform effectively (indicating a low usage divide), their lack of awareness of the underlying AI technology suggests the presence of an access divide.

In addition, negative experiences with human-AI interactions, such as those involving discrimination (outcome divide), can erode confidence in AI usage, thereby contributing to a usage divide (Hradecky et al., 2022, p. 7), or may foster a general aversion toward AI, leading to an access divide (Jain et al., 2022, p. 10). In comparison, easy access to all kinds of AI systems can lead to less questioning of the correctness of AI outputs (e.g., Uren & Edwards, 2023), resulting in a low access divide but a high outcome divide. Furthermore, Q. Wang et al. (2024) suggest that positive experiences with AI can enhance users' privacy protection skills, thereby improving outcomes (outcome divide) and, in turn, increasing AI competencies for future usage (impacting the usage divide). Those examples illustrate that the interplay between accessibility, usability, and outcomes challenges the traditional linear models that describe a linear relationship from access to usage and from usage to outcomes (e.g., Carter et al., 2020).

Moreover, our review highlights that the drivers of AI-related inequalities often span different levels of the divide. For example, while Lainjo (2023) and Oeldorf-Hirsch and Neubaum (2023) suggest that negative attitudes can hinder access to AI technologies—thus associating user attitudes with the first-level divide—(Q. Wang et al., 2024, p. 5) define the second-level divide as encompassing users' AI knowledge and skills, with the third-level divide focusing on users' attitudes. In contrast, Celik (2023), building on the work of van Dijk (2004, 2008), conceptualizes the digital divide in the context of AI technologies by distinguishing between different types of access inequalities, blending the concepts of access and usage divide. Specifically, the author identifies four types of access inequalities: motivational access, the willingness, and intention to use technologies; material access, which involves access to ICTs, the internet, and related resources; skill access, the ability to learn and effectively use AI technologies; and usage access, the actual use of these technologies. Celik (2023) empirically demonstrates that the digital divide influences AI literacy and computational thinking abilities, such as critical and problem-solving thinking, which in turn helps narrow the usage divide. This approach underscores the distinction between the drivers of inequalities—such as AI literacy and computational thinking—and the inequalities themselves.

Building upon these arguments, Oomen et al. (2023) illustrate that individuals' aversion to algorithmic tools can impede the effective implementation of AI, leading to limited outcomes. Interestingly, their case study reveals that this aversion is often not driven by the technology itself but rather by ethical concerns that arise during the implementation process, such as lack of

transparency or the inability of individuals to influence the process—drivers that are more sociotechnical in nature. This finding emphasizes the need to consider various drivers when aiming to achieve sustainable outcomes. In this example, drivers to enhance transparency, the AI outcome divide, and the third-level divide in general can be mitigated. In addition, technical AI features such as text-to-speech and speech-to-text mechanisms can lead to the “inclusion of people with disabilities in their daily life” (Schaefer et al., 2021, p. 2350), thereby, improving accessibility and contributing to sustainable outcomes. Therefore, it is crucial to distinguish between the drivers of AI inequalities in relation to the inequalities that constitute the divide itself. Understanding these distinctions allows for a more nuanced approach to addressing AI-related disparities across different levels of the divide.

To elaborate on the importance of a **multi-level analysis** of the AI divide in relation to the delta of disparities, we draw on the arguments presented by Jeon et al. (2024). The authors identify a significant AI divide between high-income and middle-income countries, distinguishing between the relative AI divide and the absolute AI divide. The relative AI divide, which refers to the gap in AI capabilities in proportional terms, has shown a more rapid convergence, which suggests that countries with fewer resources are catching up to the leading nations at a faster rate relative to their starting positions. The absolute AI divide measures the absolute differences in AI capabilities and shows a slower rate of convergence (Jeon et al., 2024), which indicates that, in absolute terms, the gap between high- and middle-income countries remains but there is considerable potential for lagging countries to catch up.

This finding underscores a critical point: countries with initially poorer access to AI resources—such as those with limited financial resources or inadequate infrastructure—can nonetheless close the gap more quickly than wealthier nations (e.g., Lainjo, 2023). This rapid convergence in relative terms can lead to better outcomes, as these countries may overcome a wider initial divide by exhibiting fewer inequalities. This suggests that the AI divide should be measured not only by absolute disparities but also by the potential for convergence and the resulting impact on inequalities. In this way, a multi-level analysis that considers relative and absolute divides provides a more nuanced understanding of AI inequalities, revealing that progress in closing the AI divide is not solely determined by the initial level of resources but also by the strategies and policies employed to bridge the gap.

4.3 The significance of the outcome divides within the sociomaterial entanglement: addressing contradictions in AI outcome inequalities

Our findings underscore the growing importance of examining the outcomes of human-AI interactions (linked to the sociomaterial entanglement in figure 4, right side) when investigating AI-related inequalities. This emphasis aligns with the concept of the third-level digital divide, as outlined in our theoretical framework (section 2.1). Specifically, our review highlights the necessity of

distinguishing between different types of outcomes as there can be contradictory effects, such as increased efficiency and decreased well-being at the same time.

For instance, implementing AI in organizations may reduce the access and usage divide among employees as they gain greater access to AI technologies and spend more time utilizing these tools (Al Shamsi et al., 2022; Jain et al., 2022). While this increased usage can lead to positive outcomes—such as enhanced trust in AI technologies and improved learning (Xing et al., 2021)—it may also result in negative consequences. These include fears of job displacement or dissatisfaction with one's work environment, driven by identity threats as AI takes over tasks traditionally performed by humans, leading to stress and diminished well-being (e.g., Mirbabaie et al., 2022; Park & Humphry, 2019; Zhang et al., 2023). These hopes and concerns also contribute to the deep-rooted discourse in business and politics about the necessity of adapting to new technologies, such as AI, in order to minimize risks and ensure the competitiveness of industries (Mansell, 2017). Consequently, while higher AI usage may mitigate performance-related inequalities, it can simultaneously exacerbate inequalities related to well-being.

In addition, Al-Fraihat et al. (2023) observed that the technology's human likeness and social presence, become significant in encouraging the adoption of smart voice assistants by online shoppers (thereby reducing the AI outcome divide of individuals). However, heightened human likeness can also induce caution and reluctance in human-AI interactions and, thus, widen the AI outcome divide between individuals. As Mou and Xu (2017, p. 437) observed that people show less open behavior and extraversion if they are aware of the AI-interaction. This may lead to increased opacity in interactions: anthropomorphic AI designs may obscure the technology's cognitive biases, while users' lack of openness can further cloud the interaction, negatively impacting outcomes (Al-Fraihat et al., 2023; Mou & Xu, 2017). For instance, addressing perceived privacy risks and increasing the explainability of AI technologies are critical steps in overcoming certain fears (e.g., García de Blanes Sebastián et al., 2022; Malodía et al., 2022).

Moreover, Hsiao and Han (2023) found that students perceive AI technologies as potentially transforming accounting professions by increasing efficiency, such as in the context of fraud detection. However, they also express concerns about ethical issues, such as the lack of communication and social interaction, threats to job security, and potential data privacy breaches (Hsiao & Han, 2023, p. 2). While AI may yield greater efficiency, these concerns limit the achievement of sustainable outcomes. Mirbabaie et al. (2022, p. 73) similarly noted that although AI can improve work-related tasks, its introduction may negatively impact employees' identification with their jobs, as AI is expected to alter workplaces and professions fundamentally. These examples illustrate that while performance-related goals often drive AI deployment in organizations, the strategies surrounding AI deployment must also address unintended consequences that pertain to sustainable outcomes, given the broader impact of AI on the workforce.

Finally, Josifidis and Supic (2024) analyzed how AI is expected to widen the productivity-pay gap in the United States further. While AI dramatically increases productivity, it does not necessarily lead to wage growth for typical workers. Stubbs (2017) also argues that the economic elite can increase their profit margins through automation. AI's efficiency in performing routine and cognitive tasks (thus reducing the AI outcome divide by increasing productivity) could disrupt traditional employment structures and negatively affect wages (thus widening the AI outcome divide by adversely impacting people's economic conditions).

These examples reveal that the linear relationship between the levels of the digital divide, as assumed in the digital divide literature (see figure 1), may explain how access and usage can lead to widening and narrowing the AI outcome divide at the same time. This discrepancy highlights the need to focus on different types of outcomes that extend beyond mere performance and efficiency in human-AI interactions (e.g., Carolus et al., 2023; Weinert et al., 2022). AI outcomes not only concern solving problems and achieving task-related goals with AI, with fewer errors (e.g., Carolus et al., 2023; Xing et al., 2021), but also distinguish between those who can and cannot reflect upon the consequences of AI utilization for individuals, organizations, societies, and the environment and identify measures to address them (e.g., Goralski & Tan, 2020).

In conclusion, contextualizing the digital divide within the framework of AI necessitates differentiating between several types of outcomes. Considering contradictory effects and enhancing the overall benefits of AI utilization is essential. This is critical for fostering a more equitable and positive impact of AI on individuals, organizations, and societies, ensuring that AI advancements contribute to immediate successes and long-term sustainability. By integrating the sustainability perspective into our conceptual framework, we can better address the complexities and challenges associated with AI-driven inequalities, thereby advancing a more nuanced and comprehensive understanding of the third-level divide in times of AI transformation.

4.4 The role of nested social subsystems: AI inequalities arise from several perspectives

To understand comprehensively how inequalities in human-AI interactions emerge, our review explores how the interdependencies within various nested perspectives of the social subsystem can contribute to disparities in human-AI interactions within an STS (see the left side of figure 4). This analysis underscores the necessity of addressing multiple drivers across different perspectives, rather than focusing on a single perspective, to mitigate inequalities.

From a *unit and team perspective*, specific groups such as developers, educators, regulators, government representatives, and managers play distinct roles, possess unique objectives, and operate within particular mindsets, all of which can influence the emergence of inequalities in human-AI interaction. For example, collaboration between developers and other teams may be essential to bridge gaps, requiring developers to reconsider certain barriers of human users as social entities when

designing AI technologies (Uren & Edwards, 2023). People from the private environment with an entrepreneurial and open mindset with respect to AI innovations, such as family members or colleagues, may act as role models, impacting AI adoption (McElheran et al., 2023). This is why all possible stakeholder groups should be integrated and involved from the beginning (Holtel, 2016). Bad social influences and peer pressure may arise within business organizations (e.g., Jain et al., 2022) that could lead to misuse and thus limit sustainable outcomes of human-AI interaction. Following Gursoy et al. (2019), social influence affects individual acceptance of AI technologies—a so-called “bandwagon effect” that is relevant when peers experience AI-based technologies (J. Wang et al., 2020). Further drivers from this perspective may include the financial wherewithal and technical equipment to drive appropriate AI initiatives (e.g., Mannuru et al., 2023; Park & Humphry, 2019), and the authority and legitimacy as a manager or developer to give instructions and set certain standards for AI deployment or design (e.g., Frick et al., 2021; Mikalef et al., 2022; Zhang et al., 2023).

From an *institutional and organizational perspective*, organizational culture—including change capacity, risk tolerance, interdepartmental collaboration, and leadership style—as well as internal infrastructure, processes, and governance mechanisms significantly affect whether organizations can successfully utilize AI technologies. For instance, organizations with greater financial resources are better positioned to invest in AI, a factor often linked to leadership style and organizational type. Startups, for example, tend to have younger, more educated, and more experienced leaders who bring new ideas to market and contribute to the community (McElheran et al., 2023). High initial capitalization also plays a crucial role in the adoption of AI within organizations (McElheran et al., 2023). Also, inequalities may arise from inconsistent leadership values (e.g., Bérubé et al., 2021). Thus, an organization’s financial resources, type, and leadership style impact whether or not it can successfully deploy AI technologies.

Q. Wang et al. (2024, p. 1012134) further argued that in an “organizational context, regulatory quality, government effectiveness, government expenditure, rule of law, and corruption control play pivotal roles.” The authors add that these drivers not only enhance operational efficiency but also ensure that ethical considerations are upheld in the deployment of AI. Accordingly, developing guidelines and handbooks to ensure that organizational processes are ethically sound is essential (Aydin et al., 2023). Without robust institutional governance guidelines, unintended consequences and negative impacts on individuals, organizations, or societies can arise (e.g., Mirbabaie et al., 2022). External institutions, such as governmental bodies that enforce anti-discrimination policies and ensure compliance with legal obligations, play a vital role in this regard. Aydin et al. (2023, p. 10). Emphasized that “organisations should register for being monitored by interdependent institutions to show that they consider ethical and inclusionary approaches to employees and

applicants.” This highlights the importance of focusing on sustainable outcomes in AI adoption projects, which intersect with legal frameworks and laws from the societal perspective.

Corporate culture can also influence inequalities in human-AI interaction. For instance, a digital workplace culture that includes “a data-driven perspective that adopts a combination of work at the office and from home” (Aydin et al., 2023), can positively affect data management practices in human-AI interactions, thereby helping to reduce the AI divide. Organizational culture also encompasses an institution’s capacity for change management, risk tolerance, and interdepartmental collaboration (Mikalef et al., 2022). In addition, Hsiao and Han (2023, p. 2) argued that companies must establish an infrastructure that fosters the development of in-demand skills, addressing the “gap between the number of positions demanding [data analytics] and AI skills and the number of qualified candidates.” Thus, human resource management (HRM) processes aimed at reducing skill disparities within the workforce also influence inequalities from an individual perspective.

Collectively, these drivers are critical for the successful implementation of AI in organizations (Mikalef et al., 2022). They also contribute to sustainable outcomes. For example, Aydin et al. (2023, p. 2) demonstrate that “gender-based (in)equalities are shaped and reproduced within Industry 5.0 by considering the role of ethical issues, machine ethics and the co-creation of values amongst human resources and software development teams in supplying human capital.” Based on their findings, the authors suggest that value co-creation through collaboration between human resources and software development teams, particularly in the design of machines for business processes, should be grounded in digital workplace culture, norms, ethical codes, and principles. This emphasizes that corporate culture influences inequalities among groups such as developers or HR personnel at the unit level (*unit perspective*), which in turn affects the potential for achieving sustainable outcomes in human-AI interactions.

From an *industrial and market perspective*, our review highlights several key drivers that differentiate industries with a higher potential for sustainably utilizing AI technologies. These drivers include varying levels of security restrictions (e.g., Hradecky et al., 2022), customer hesitancy in adopting new technologies (e.g., Malodia et al., 2022), and the potential of AI and robotics to enhance the value chain and transform production methods, as seen in intelligent farming and manufacturing (e.g., Gao, 2024). Notably, sectors such as manufacturing and education have the lowest rates of AI adoption, while healthcare, financial services, and management and administration exhibit the highest rates (McElheran et al., 2023). In particular, an increased use of AI was observed in companies characterized by dynamism and well-educated, experienced owners (McElheran et al., 2023). AI is frequently used in high-growth startups that utilize the latest product and process innovations (McElheran et al., 2023). According to Josifidis and Supic (2024), the impact of technological advancements on wages and productivity varies across industries. For instance, the manufacturing and retail sectors have experienced significant automation, leading to wage stagnation

for workers in routine jobs. The authors suggest that this productivity-pay gap is a major driver of income inequality, particularly in the context of disruptive changes in the United States markets.

Hence, the divide between industries and markets in achieving sustainable AI outcomes is influenced by security restrictions, customer adoption rates, and the potential for AI to transform value chains. Industries such as healthcare and financial services are leading in AI adoption, while sectors such as manufacturing and education lag behind. This uneven adoption contributes to disparities in productivity and income, enforcing inequalities for organizations and individuals (corresponding to the other perspectives) and highlighting the need for tailored strategies to address these challenges and ensure equitable AI-driven progress across all sectors.

From a *societal, geographical, and environmental perspective*, certain countries, nations, and regions may be better positioned than others to harness AI benefits. Inequalities can arise due to inadequate infrastructure and the absence of digital connectivity (e.g., Bérubé et al., 2021), financial capacities to allocate resources for public sector innovations (e.g., Q. Wang et al., 2024), a non-open technology societal culture (e.g., Kamoonpuri & Sengar, 2023), misconceptions about the nature of intelligence of AI technologies compared to human beings (e.g., Gignac & Szodorai, 2024), brain drain effects on governmental AI readiness (e.g., Iuga & Socol, 2024), disparities between high-income and middle-income countries (e.g., Jeon et al., 2024), and global AI trends and practices (e.g., Q. Wang et al., 2024).

The work of Josifidis and Supic (2024) examines the impact of AI on the productivity-pay gap in the United States, finding that AI is likely to further widen this gap. They argue that heterodox economics, which views the economy as a social provisioning process, offers valuable insights into understanding and mitigating the negative effects of AI on income distribution and the labor market. Their proposed solutions include government intervention, universal basic income, and public employment schemes. The authors caution that without regulation, AI could exacerbate existing inequalities and so recommend policy interventions such as shorter workweeks, expanded social safety nets, and taxation of AI technologies to fund public services or a universal basic income. They conclude that AI's impact on the productivity-pay gap and income distribution could be significant, necessitating proactive measures to prevent increasing inequality. From a different perspective, Yuan et al. (2024) found that AI has the potential to reduce the wage gap in certain industries, particularly in technology-intensive sectors. This highlights the importance of analyzing AI's effects from multiple perspectives to understand whether it will exacerbate or alleviate inequalities.

Gao (2024, p. 1) examined regional economic inequalities between Chinese cities, noting that in cities with high robotics development, regional inequality is greater, which is rooted in agglomeration as well as differentiation effects. Robotics, thus, promotes income, investment, regional consumption, and public services, driving economic growth (Gao, 2024). However, unequal

innovation in urban and rural areas and in developed and less-developed regions exacerbates regional inequalities, as differentiation effects are the main cause of the growing imbalance (Gao, 2024). The author emphasized the need for governments to “strike a balance between ‘efficiency’ and ‘fairness’” (p. 30), highlighting that economic objectives are overly concentrated in China’s large urban areas, resulting in negative externalities, while suburban areas remain underserved. Therefore, addressing inequalities that hinder sustainable AI utilization across countries is crucial.

This view is reinforced by Lainjo (2023, p. 4967), who argues that “ethical AI development must consider society’s diverse demands and beliefs. AI must be human-centered to protect vulnerable groups and build public trust. Studying AI-related social dynamics and inequities is essential for creating a technologically advanced, egalitarian, and morally responsible future. By addressing biases, discrimination, the digital divide, and ethical concerns, we can maximize AI’s potential while ensuring its benefits are shared by all.” These examples demonstrate that addressing inequalities of sustainable outcomes of human-AI interaction from a societal perspective also impacts the other perspectives, such as developers at the unit level receiving guidelines for inclusive development (e.g., Mannuru et al., 2023) or employees at the organizational level benefiting from considerations to avoid income inequalities.

When considering geographical differences, McElheran et al. (2023, p. 398) determined that “large metro areas [in the United States], such as Nashville, San Antonio, Las Vegas, New Orleans, San Diego, and Tampa, also show high concentrations of AI use. The concentration of AI in already-leading locations suggests an ‘AI divide’ that may be further reinforced if new AI-using, high-growth startups continue to be attracted to locations with already-high-AI activity.” Daepp and Counts (2024) identified geographical patterns in user searches for ChatGPT between the West Coast and Gulf states and showed that countries with the highest rates of search are more urbanized and educated. This points to further interdependencies between the industrial and the geographical perspective that must be considered when reducing disparities in creating sustainable outcomes of human-AI interaction.

Considering environmental differences in their study, Zhao et al. (2024) analyzed data from 74 countries between 2000 and 2019 to examine the asymmetric impact of AI on carbon inequalities and discovered a negative causal relationship, indicating that AI helps reduce carbon inequalities. Specifically, the authors observed that “[w]ith the increase of the quantile of carbon inequality, artificial intelligence exerts a more remarkable inhibiting effect on carbon inequality, which implies that when the level of carbon inequality is more severe, the implementation of artificial intelligence proves to be a more potent tool for narrowing the disparity of emissions.” (Zhao et al., 2024, p. 140817). Aydin et al. (2023) highlight the potential to reduce CO₂ emissions through automation of manufacturing and logistics, and Goralski and Tan (2020) the potential to save water resources. However, large-scale AI models and the widespread use of AI applications also have high levels of

energy consumption (e.g., Mannuru et al., 2023). Hence, the ability of countries to mitigate environmental inequalities linked to saving or using resources through AI becomes more significant as the severity of these disparities increases.

The work of Mannuru et al. (2023) involves different perspectives when examining inequalities in the context of generative AI. Following the authors, from an *individual perspective* there can be inequalities in reflecting how AI hinders interpersonal communication, raises access and equity concerns, and devalues traditional research methods or how data can be sourced, structured, and interpreted (Mannuru et al., 2023). From an *organizational perspective*, there can be inequalities in achieving efficiency, productivity, customer experiences, compensating for gender inequities, ethnic and wealth barriers, overreliance on AI, job displacements, ethical concerns, and AI-related bias and discrimination (Mannuru et al., 2023). From a *societal perspective*, there can be benefits through better climate forecasting, carbon storage, calculation of energy performance ratings, patient care, less waste of energy and water, reduction of privacy concerns, and regulatory complexity. The different outcomes are thus interrelated. For instance, inefficient use of AI could lead to resource waste, exacerbating the negative consequences of AI (Mannuru et al., 2023). Conversely, AI can help to achieve sustainability goals by conserving precious resources, such as offering smarter management of water, including the transportation and reuse of water (Goralski & Tan, 2020).

In conclusion, a multi-level perspective is essential for understanding and addressing inequalities in human-AI interactions. By considering the interdependencies across various perspectives within nested social subsystems—from individual and team dynamics to organizational culture and institutional frameworks to societal forces—this analysis demonstrates that a holistic approach is necessary to mitigate disparities. Addressing these drivers of inequalities collectively, rather than in isolation, is crucial for ensuring that AI technologies are implemented in ways that are sustainable, ultimately leading to more equitable outcomes.

5. Discussion

In this section, we discuss the results of our empirical investigation, offer theoretical and practical implications, and address the limitations of our research.

5.1 Theoretical implications: Rethinking the AI divide by using SST and integrating a sustainability development perspective

Our conceptualization of the AI divide builds upon Leonardi's (2012) framework of sociotechnical systems. SST is useful for examining how sociomaterial practices between humans and AI are shaped by the arrangement of human-related objects (e.g., management) and non-human objects (e.g., governance) within the nested social subsystem. By applying this theory from a sustainable development perspective, we highlight circular interdependencies between the divide levels, examine

the interplay of inequalities emerging from different perspectives within the nested social subsystem, and explore ambivalent outcomes of human-AI interactions. These insights lead to three theoretical implications that extend previous divide models and respond to the call of Q. Wang et al. (2024, p. 15), who “ recommend future studies to further examine whether the model in the context of the AI divide will differ from conventional digital divide studies in the future.”

Regarding the *circularity of the divide levels*, our review highlights that all divide levels are interrelated in some way. For instance, in underprivileged regions, education is a great predictor of AI usage and creates interest in AI technologies and knowledge for future interactions (Daepf & Counts, 2024; W. Li et al., 2024). Thus, education not only fosters inclusivity as a sustainable outcome but also enhances learning. In turn, these improvements create better opportunities for individuals to engage with additional technologies, thereby reducing access inequalities. Conversely, in the healthcare context, AI tools that are not accurately trained on diverse datasets tend to be less effective for certain populations (e.g., minorities) (e.g., Lainjo, 2023). This outcome divide can exacerbate health disparities, as these populations may increasingly avoid AI-based diagnostic or treatment tools in future medical scenarios, thus deepening access inequalities. This example illustrates how inequalities at different levels of the divide can themselves become drivers of new forms of inequality. Future work could elaborate on this circularity.

Regarding the *interdependencies of the nested perspectives*, SST also offers a comprehensive framework for understanding potential barriers to inequalities by incorporating nested perspectives within the social subsystem—including unit/team, institutional/organizational, industrial/market, societal/environmental perspective—and examining their interdependencies. For instance, our findings suggest governance mechanisms may arise from various perspectives—for example, industry-specific policies such as retail chatbot regulation that safeguards customer data, or broader societal regulation that enhances ethical AI capabilities of organizations (e.g., Kamoopuri & Sengar, 2023; Mikalef et al., 2022; Zhang et al., 2023). The literature also highlights that governance appears within different structures, such as on a technical or organizational level (e.g., Shneiderman, 2020). While specific regulations such as the European AI Act can guide the governance mechanism of industries, organizations, and developers teams from the top down, there are also bottom-up AI initiatives that inform future AI regulation (e.g., Aslan et al., 2022; Westerstrand, 2023). Future research could investigate whether these interdependencies among the nested perspectives follow a linear progression from the societal level down to the industrial and organizational levels or whether they exhibit more complex relationships.

Regarding the *ambivalent outcomes in human-AI interactions*, our findings underscore the importance of considering potential conflicts when examining inequalities of human-AI interaction from a sustainable development perspective that extends beyond performance-related issues. For instance, while regulation and AI governance across all levels generally support sustainable

outcomes, overregulation could potentially stifle AI innovation, thereby hindering the creation of beneficial outcomes (Allen, 2019; Han et al., 2021). Moreover, individuals who become accustomed to using AI may be less likely to assess their own usage behavior critically, especially with respect to unintended consequences (Carolus et al., 2023). This form of overtrust in AI could undermine sustainable outcomes while supporting efficiency outcomes: users may develop trust and increase efficiency, especially when AI systems behave in a human-like manner (Klumpp, 2017). However, this may simultaneously impede the development of realistic narratives regarding what AI can or cannot do (Diederich et al., 2022), as individuals may forget how to distinguish between the emotions they experience when interacting with real people and AI (Al-Fraihat et al., 2023). This can lead to overtrust, which is especially critical in unforeseen situations, such as emergency situations (e.g., Wagner, 2019). In an experiment in which participants interact with robots in an emergency evacuation scenario, for instance, Robinette et al. (2017) determined that the majority of people tend to rely on the robot, which demonstrates overtrust.

These dynamics suggest that drivers within a sociotechnical system can foster inequalities and qualities simultaneously, leading to certain ambivalences. Therefore, associating sustainable interactions with long-term success is not an absolute rule; tensions may arise between achieving sustainability and meeting other success criteria, particularly in the short term. These complexities highlight the need for a balanced approach and further empirical investigations that consider both immediate performance and efficiency outcomes and long-term sustainability in the design and implementation of AI systems.

To summarize, our conceptualization of the AI divide, grounded in Leonardi's (2012) sociotechnical systems framework, provides three key theoretical implications that extend previous models of the digital divide. First, the circular interdependencies between different levels of the divide demonstrate how inequalities can propagate and exacerbate one another, suggesting that sustainable development initiatives must account for these complex interactions. Second, the interdependencies within the nested perspectives of the social subsystem reveal that drivers of inequalities operate across multiple levels, from unit to organizational to societal, and these may interact in non-linear and complex ways. This challenges the conventional top-down approach to overcome certain drivers of inequalities and calls for more nuanced studies into these relationships. Third, the ambivalence between different types of outcomes of human-AI interactions highlights potential conflicts, especially where short-term success may undermine long-term sustainability. Our conceptualization is thus consistent with the idea of HCAI research, putting human values into the core by highlighting the relevance of a sustainability perspective besides performance goals (Andersen et al., 2022; van Berkel et al., 2022).

5.2 Practical implications: managing AI inequalities in alignment with the SDGs

The review identifies several strategies for addressing AI-induced inequalities, contributing to multiple SDGs of the United Nations (2020). Various stakeholders within the nested social subsystem, including policymakers, leaders, educators, and developers, can play a pivotal role in bridging the AI divide through appropriate management strategies aligned with the SDGs.

From a *unit perspective* of developers, fostering responsible AI begins with enhancing their own literacy regarding the potential for AI to exacerbate social inequalities (Oeldorf-Hirsch & Neubaum, 2023). A crucial element of designing responsible and inclusive AI involves collaboration between developers and prospective AI user groups. Such partnerships can help identify and mitigate potential biases and barriers early in the development process, thereby reducing obstacles before novel AI products are released (Mannuru et al., 2023; Uren & Edwards, 2023). These efforts not only promote greater inclusivity of AI technologies, aligning with SDG#16 (peace, justice, and strong institutions) but also advancing gender equality, as put forth in SDG#5 (gender equality).

From an *organizational perspective*, managers should implement an AI strategy that establishes clear guidelines for AI usage and provides employees with training on ethical frameworks, ensuring they are aware of their rights (Mikalef et al., 2022). Such initiatives contribute to SDG#4 (quality education) by promoting equitable access to quality education and lifelong learning opportunities, as well as SDG#8 (decent work and economic growth) by fostering productive employment. Moreover, supportive leadership can help bridge the gap between AI-proficient employees and those less familiar with the technology (Frick et al., 2021). This might further enhance inclusivity within organizations (SDG#16; United Nations, 2020).

From a *market perspective*, AI providers can integrate specific nudges within their marketing strategies to encourage sustainable online shopping (Malodia et al., 2022; Stolz et al., 2024). These approaches align with SDG#12 (responsible consumption and production). In addition, by reducing barriers to accessing AI technologies across diverse user groups, providers can both narrow the digital divide and empower a broader range of individuals to participate within society (SDG#5) and support sustainable innovations (SDG#9; United Nations, 2020).

From a *societal perspective*, educational strategies must be reexamined to include training that enhances student awareness of both the benefits and potential threats of AI, fostering critical reflection skills (Zhang et al., 2023). Such curriculum updates not only improve the quality of education (SDG#4) but also contribute to student well-being by increasing their awareness of the mental health risks of using AI (SDG#3; United Nations, 2020). Furthermore, policymakers should work toward aligning guidelines and establishing minimum standards for AI usage, as a lack of ethical oversight may exacerbate social disparities and exclude individuals from equal participation

in society (Choudrie et al., 2022). This approach can strengthen global partnerships (SDG#17) and inclusive institutions at all levels (SDG#16; United Nations, 2020).

Table 1 is an overview of the recommendations for managing AI in alignment with the SDGs.

Perspective	Recommendations for Managing AI Inequalities	SDGs of the United Nations (2020)
Unit (e.g., developers)	Integrate ethical design principles into AI systems from the outset by collaborating with prospective AI user groups. This includes considering the social and ethical implications of AI technologies, such as fairness, transparency, and accountability, to prevent unintended consequences through certain design biases that may exacerbate inequalities.	SDG#5 (gender equality), SDG#16 (peace, justice, and strong institutions)
	Establish a continuous feedback loop with a diverse range of AI users to monitor the real-world impact of AI systems. This ongoing engagement will allow developers to identify and address any emerging inequalities or biases in AI applications, ensuring the systems evolve in a way that promotes inclusivity and fairness.	
Organizational (e.g., managers)	Develop a comprehensive AI governance strategy within the organization, guided by an ethical committee that includes stakeholders from various sectors such as society, industry, and other relevant organizations. This strategy should be communicated effectively and translated into clear, comprehensible guidelines for employees that detail how AI systems are used within the organization and the underlying reasons for these practices.	SDG#4 (quality education quality), SDG#8 (decent work and economic growth), SDG#16 (peace, justice, and strong institutions)
	Create compelling role models for AI implementation and cultivate a leadership culture that is supportive of AI-related issues faced by employees. This can be achieved by providing leadership training that sensitizes leaders to the importance of sustainable human-AI interactions. In addition, communicate to employees within the organization who possess specific AI expertise and can be consulted for assistance, for instance, by providing an internal “map” of expert users.	
Industrial (e.g., AI providers)	Conduct customer surveys to identify potential barriers that arise from missing or inadequate touchpoints with AI throughout a typical customer journey. Once identified, take proactive steps to remove these barriers, ensuring a smoother and more inclusive experience for all customers.	SDG#5 (gender equality), SDG#9 (sustainable innovation), SDG#12 (responsible consumption)
	Make AI technologies easily accessible and practical for end-users. This requires maintaining a high standard of system quality, which is essential for effectively implementing the established principles of AI governance.	
Societal (e.g., policymakers, educators)	Reevaluate AI curriculum design to ensure it equips individuals with the necessary skills for achieving sustainable AI outcomes. This may involve challenging current educational standards to determine whether they adequately address the abilities required for effective AI usage—for example, through participation in teaching and tutor evaluations.	SDG#3 (well-being), SDG#4 (quality education quality)
	Develop clearly understandable and comprehensible AI regulations through a dialogue involving researchers, educators, managers, and AI users. These regulations should account for the evolving landscape of AI law and reflect current legal frameworks, such as data protection laws, that impact AI governance.	

Table 1 Recommendations for managing AI inequalities related to the SDGs.

5.3 Limitations and future work

Our study has several limitations that highlight potential avenues for future research.

First, our review is grounded in the specific perspective of SST, which inherently limits our synthesis of the examined literature to the SST framework. Furthermore, our study lacks empirical validation. Future research could delve into specific relationships identified in our conceptual framework that have yet to be empirically tested, and may also reconsider theoretical approaches to structuring inequalities. For example, future work could develop a measurement instrument for the AI divide, enabling an empirical analysis to explore the circular relationships between different levels of the divide.

Second, while in our conceptualization we acknowledge the influence of various perspectives within the social subsystem (e.g., societal, industrial, organizational) on barriers to human-AI interaction, a more in-depth exploration of how these perspectives interact is necessary. Our review favors breadth over depth. Future research could examine these perspectives more thoroughly, perhaps by investigating interdependencies through qualitative analyses, such as expert interviews with representatives from different sectors (e.g., governmental and market). This approach could enhance our understanding of the factors that widen or narrow the AI divide within various micro- and macro-perspectives.

Third, although we aimed to ensure systematicity and transparency in our review process by adhering to the steps outlined by Templier and Paré (2018), our main analysis and discussion are based primarily on intensive discussions among the three coders. In addition, the literature we reviewed was subject to several screening criteria. Given the increasing discourse on AI inequalities, it would be beneficial for future research to include non-peer-reviewed work and research-in-progress articles to capture a broader range of perspectives.

5. Conclusion

In this work, we use a theoretical literature review to understand how and why inequalities arise when humans interact with AI. We conceptualize the AI divide as a circular model and follow SST to demonstrate how inequalities can arise that may hinder sustainable outcomes of human-AI interactions. We assume that several outcomes are relevant for an equal distribution of AI's benefits among individuals, organizations, and societies and, thus, for supporting the United Nations SDGs. We further demonstrate that different components of the sociotechnical system can widen and narrow the AI divide at the same time—indicating that there can be inequalities that somehow act in contrasting or supporting ways. We hope our findings provide valuable implications for future research and practitioners.

References

- Al Shamsi, J. H., Al-Emran, M., & Shaalan, K. (2022). Understanding key drivers affecting students' use of artificial intelligence-based voice assistants. *Education and Information Technologies*, 27(6), 8071–8091. <https://doi.org/10.1007/s10639-022-10947-3>
- Al-Fraihat, D., Alzaidi, M., & Joy, M. (2023). Why do consumers adopt smart voice assistants for shopping purposes? A perspective from complexity theory. *Intelligent Systems with Applications*, 18, 200230. <https://doi.org/10.1016/j.iswa.2023.200230>
- Allen, T. C. (2019). Regulating Artificial Intelligence for a Successful Pathology Future. *Archives of Pathology & Laboratory Medicine*, 143(10), 1175–1179. <https://doi.org/10.5858/arpa.2019-0229-ED>
- Alsheibani, S., Cheung, Y., & Messom, C. (2019). Factors inhibiting the adoption of artificial intelligence at organizational-level: A preliminary investigation. *AMCIS 2019 Proceedings*, 2. <https://research.monash.edu/en/publications/factors-inhibiting-the-adoption-of-artificial-intelligence-at-org>
- Alsheibani, S., Messom, C., Cheung, Y., & Alhosni, M. (2020). Artificial Intelligence Beyond the Hype: Exploring the Organisation Adoption Factors. *ACIS 2020 Proceedings*. <https://aisel.aisnet.org/acis2020/33>
- Andersen, R., Mørch, A. I., & Litherland, K. T. (2022). Collaborative learning with block-based programming: Investigating human-centered artificial intelligence in education. *Behaviour & Information Technology*, 41(9), 1830–1847. <https://doi.org/10.1080/0144929X.2022.2083981>
- Aslan, A., Greve, M., & Lembcke, T.-B. (2022). *Let's Do Our Bit: How Information Systems Research Can Contribute to Ethical Artificial Intelligence*. *AMCIS Proceedings*. https://www.researchgate.net/profile/aycan-aslan/publication/369480408_lets_do_our_bit_how_information_systems_research_can_contribute_to_ethical_artificial_intelligence/links/641d70bb66f8522c38cd02ec/let-s-do-our-bit-how-information-systems-research-can-contribute-to-ethical-artificial-intelligence.pdf
- Avgerou, C. (2019). Contextual Explanation: Alternative Approaches and Persistent Challenges. *Management Information Systems Quarterly*, 43(3), 977–1006. <https://aisel.aisnet.org/misq/vol43/iss3/15>

- Aydin, E., Rahman, M., & Ozeren, E. (2023). Does Industry 5.0 Reproduce Gender (In)equalities at Organisations? Understanding the Interaction of Human Resources and Software Development Teams in Supplying Human Capitals. *Information Systems Frontiers*, 1–15. <https://doi.org/10.1007/s10796-023-10450-1>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484*.
- Baker, J. (2011). The Technology–Organization–Environment Framework. In Y. K. Dwivedi, M. R. Wade, & S. L. Schneberger (Eds.), *Integrated Series in Information Systems Ser: v.28. Information Systems Theory* (Vol. 28, pp. 231–245). Springer. https://doi.org/10.1007/978-1-4419-6108-2_12
- Bérubé, M., Giannelia, T., & Vial, G. (2021). Barriers to the Implementation of AI in Organizations: Findings from a Delphi Study. In T. Bui (Ed.), *Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 54th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2021.805>
- Blank, G., & Lutz, C. (2018). Benefits and harms from Internet use: A differentiated analysis of Great Britain. *New Media & Society*, 20(2), 618–640. <https://doi.org/10.1177/1461444816667135>
- Broadbent, R., & Papadopoulos, T. (2013). Bridging the digital divide – an Australian story. *Behaviour & Information Technology*, 32(1), 4–13. <https://doi.org/10.1080/0144929X.2011.572186>
- Brock, J. K.-U., & Wangenheim, F. von (2019). Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. *California Management Review*, 61(4), 110–134.
- Bughin, J., & van Zeebroeck, N. (2018, July 26). Artificial intelligence: Why a digital base is critical. *McKinsey & Company*. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/artificial-intelligence-why-a-digital-base-is-critical>
- Carolus, A., Augustin, Y., Markus, A., & Wienrich, C. (2023). Digital interaction literacy model – Conceptualizing competencies for literate interactions with voice-based AI systems. *Computers and Education: Artificial Intelligence*, 4.

- Carter, L., Liu, D., & Cantrell, C. (2020). Exploring the Intersection of the Digital Divide and Artificial Intelligence: A Hermeneutic Literature Review. *AIS Transactions on Human-Computer Interaction*, 12(4), 253–275. <https://doi.org/10.17705/1thci.00138>
- Celik, I. (2023). Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption. *Telematics and Informatics*, 83, 102026. <https://doi.org/10.1016/j.tele.2023.102026>
- Chang, Y. S., Jeon, S., Jo, S. J., & Kim, Y. E. (2022). *Speed of Catch-up and Convergence of the Artificial Intelligence Divide: Robotic and Patents*. <https://doi.org/10.2139/ssrn.4140776>
- Choudrie, J., Zamani, E., & Obuekwe, C. (2022). Bridging the Digital Divide in Ethnic Minority Older Adults: an Organisational Qualitative Study. *Information Systems Frontiers*, 24(4), 1355–1375.
- Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, 33(1), 100899. <https://doi.org/10.1016/j.hrmr.2022.100899>
- Cohen, J. (1960). *A Coefficient of Agreement for Nominal Scales* (Vol. 20). <https://psycnet.apa.org/record/1960-06759-001> <https://doi.org/10.1177/001316446002000104>
- Daepp, M. I. G., & Counts, S. (2024, April 18). *The Emerging AI Divide in the United States*. <http://arxiv.org/pdf/2404.11988v2>
- Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, 1(2), 73–80. <https://doi.org/10.1080/2573234X.2018.1543535>
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3), 319. <https://doi.org/10.2307/249008>
- Diederich, S., Brendel, A. B [Alfred Benedikt], Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1), 96–138.
- Diel, S., Höger, C., & Schick, D. (2021). Alexa – Welcome to the Family! IT Identity's Mediating Role on Social Presence and Deep Use of In-Home Voice Assistants. *ICIS 2021 Proceedings*. https://aisel.aisnet.org/icis2021/soc_impact/soc_impact/5
- Dwivedi, Y. K., Kshetri, N., Hughes, L., ..., & Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative

- conversational AI for research, practice and policy. *International Journal of Information Management*, 71.
- Endacott, C. G., & Leonardi, P. M. (2022). Artificial Intelligence and Impression Management: Consequences of Autonomous Conversational Agents Communicating on One's Behalf. *Human Communication Research*, 48(3), 462–490. <https://doi.org/10.1093/hcr/hqac009>
- Frick, N. R. J., Mirbabaie, M., Stieglitz, S., & Salomon, J. (2021). Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises. *Journal of Decision Systems*, 30(2-3), 235–258.
- Gao, Z. (2024). Shadow of Technology: Robotics Development and Regional Economic Inequality. *Journal of the Knowledge Economy*, 1–34. <https://doi.org/10.1007/s13132-024-01856-6>
- García de Blanes Sebastián, M., Sarmiento Guede, J. R., & Antonovica, A. (2022). Application and extension of the UTAUT2 model for determining behavioral intention factors in use of the artificial intelligence virtual assistants. *Frontiers in Psychology*, 13, 993935. <https://doi.org/10.3389/fpsyg.2022.993935>
- Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence*, 104, 101832. <https://doi.org/10.1016/j.intell.2024.101832>
- Goldenthal, E., Park, J., Liu, S. X., Mieczkowski, H., & Hancock, J. T. (2021). Not All AI are Equal: Exploring the Accessibility of AI-Mediated Communication Technology. *Computers in Human Behavior*, 125, 106975. <https://doi.org/10.1016/j.chb.2021.106975>
- Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1), 100330. <https://doi.org/10.1016/j.ijme.2019.100330>
- Grundspenkis, J. (2007). Agent based approach for organization and personal knowledge modelling: Knowledge management perspective. *Journal of Intelligent Manufacturing*, 18(4), 451–457. <https://doi.org/10.1007/s10845-007-0052-6>
- Gupta, S [Shivam], Kamboj, S., & Bag, S. (2023). Role of Risks in the Development of Responsible Artificial Intelligence in the Digital Healthcare Domain. *Information Systems Frontiers*, 25(6), 2257–2274. <https://doi.org/10.1007/s10796-021-10174-0>

- Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, 49, 157–169. <https://doi.org/10.1016/j.ijinfomgt.2019.03.008>
- Han, T. A., Pereira, L. M., Lenaerts, T., & Santos, F. C. (2021). Mediating artificial intelligence developments through negative and positive incentives. *PloS One*, 16(1), e0244592. <https://doi.org/10.1371/journal.pone.0244592>
- Hansen, H. F., Lillesund, E., & Mikalef, P [Patrick]. (2022). *Examining AI adoption through a maturity model: A Case Study on Norwegian Organizations*. The 5th Middle East & North Africa Conference for Information System (MENACIS 2022).
- Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3), 101772.
- Holtel, S. (2016). Artificial Intelligence Creates a Wicked Problem for the Enterprise. *Procedia Computer Science*, 99, 171–180. <https://doi.org/10.1016/j.procs.2016.09.109>
- Hradecky, D., Kennell, J., Cai, W., & Davidson, R. (2022). Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe. *International Journal of Information Management*, 65, 102497. <https://doi.org/10.1016/j.ijinfomgt.2022.102497>
- Hsiao, D., & Han, L. (2023). THE IMPACT OF DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE ON THE FUTURE ACCOUNTING PROFESSION: PERSPECTIVES FROM ACCOUNTING STUDENTS. *Journal of Theoretical Accounting Research*.
- Iuga, I. C., & Socol, A. (2024). GOVERNMENT ARTIFICIAL INTELLIGENCE READINESS AND BRAIN DRAIN: INFLUENCING FACTORS AND SPATIAL EFFECTS IN THE EUROPEAN UNION MEMBER STATES. *Journal of Business Economics and Management*, 25(2), 268–296. <https://doi.org/10.3846/jbem.2024.21136>
- Jain, R., Garg, N., & Khera, S. N. (2022). Adoption of AI-Enabled Tools in Social Development Organizations in India: An Extension of UTAUT Model. *Frontiers in Psychology*, 13, 893691. <https://doi.org/10.3389/fpsyg.2022.893691>
- Jeon, S., Chang, Y. S., Jo, S. J., Madukuand, T., & Kim, Y. E. (2024). Speed of Catch-Up and Convergence of the Artificial Intelligence Divide: AI Investment, Robotic, Start-Ups, and Patents. *Journal of Global Information Technology Management*, 27(1), 63–85. <https://doi.org/10.1080/1097198X.2023.2297636>

- Jöhnk, J., Weißert, M., & Wyrski, K. (2021). Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors. *Business & Information Systems Engineering*, 63(1), 5–20.
- Josifidis, K., & Supic, N. (2024). Artificial Intelligence and the Productivity-Pay Gap in the USA: Industrial Insights and the Revival of Heterodox Ideas. *Journal of Economic Issues*, 58(2), 540–548. <https://doi.org/10.1080/00213624.2024.2344434>
- Kamoonpuri, S. Z., & Sengar, A. (2023). Hi, May AI help you? An analysis of the barriers impeding the implementation and use of artificial intelligence-enabled virtual assistants in retail. *Journal of Retailing and Consumer Services*, 72, 103258. <https://doi.org/10.1016/j.jretconser.2023.103258>
- Kar, A. K., Choudhary, S. K., & Singh, V. K. (2022). How can artificial intelligence impact sustainability: A systematic literature review. *Journal of Cleaner Production*, 376, 134120. <https://doi.org/10.1016/j.jclepro.2022.134120>
- Kelan, E. K. (2024). Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. *Human Resource Management Journal*, 34(3), 694–707. <https://doi.org/10.1111/1748-8583.12511>
- Keskin, T., & Vermeulen, F. (2024). *Proceedings of the 57th Annual Hawaii International Conference on System Sciences: Hilton Hawaiian Village Waikiki Beach Resort, January 3-6, 2024*. Department of IT Management Shidler College of Business University of Hawaii. <https://scholarspace.manoa.hawaii.edu/items/179cefa9-1bd3-4440-a68b-4f5e5ae530f0>
<https://doi.org/107396>
- Kitsara, I. (2022). Artificial Intelligence and the Digital Divide: From an Innovation Perspective. *Progress in IS*, 245–265. https://ideas.repec.org/h/spr/prochp/978-3-030-90192-9_12.html
- Klumpp, M. (2017). Artificial Divide: The New Challenge of Human-Artificial Performance in Logistics. In H. Proff & T. M. Fojcik (Eds.), *SpringerLink Bücher. Innovative Produkte und Dienstleistungen in der Mobilität: Technische und betriebswirtschaftliche Aspekte* (pp. 583–593). Springer Gabler. https://doi.org/10.1007/978-3-658-18613-5_37
- Klumpp, M., & Zijm, H. (2019). Logistics Innovation and Social Sustainability: How to Prevent an Artificial Divide in Human–Computer Interaction. *Journal of Business Logistics*, 40(3), 265–278. <https://doi.org/10.1111/jbl.12198>
- Kolbjørnsrud, V., Amico, R., & Thomas, R. J. (2017). Partnering with AI: how organizations can win over skeptical managers. *Strategy & Leadership*, 45(1), 37–43.

- Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In *AHFE International, The Human Side of Service Engineering*. AHFE International. <https://doi.org/10.54941/ahfe1002565>
- Lainjo, B. (2023). *The Global Social Dynamics and Inequalities of Artificial Intelligence*. <https://doi.org/10.13140/RG.2.2.20078.79689>
- Lainjo, B. (2024). The implications of Artificial Intelligence on international development management. *Journal of Autonomous Intelligence*, 7(5), 1258. <https://doi.org/10.32629/jai.v7i5.1258>
- Laut, P., Dumbach, P., & Eskofier, B. (2021). Integration of Artificial Intelligence in the Organizational Adoption – A Configurational Perspective. *ICIS 2021 Proceedings*. https://aisel.aisnet.org/icis2021/is_implement/is_implement/2
- Legris, P., Ingham, J., & Colletette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management*, 40(3), 191–204. [https://doi.org/10.1016/S0378-7206\(01\)00143-4](https://doi.org/10.1016/S0378-7206(01)00143-4)
- Leonardi, P. M. (2012). Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How are They Related? Do We Need Them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and Organizing: Social Interaction in a Technological World* (pp. 25–48). Oxford University Press. <https://doi.org/10.2139/ssrn.2129878>
- Leslie, D. (2019). *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. <https://doi.org/10.2139/ssrn.3403301>
- Li, M., & Suh, A. (2022). Anthropomorphism in AI-enabled technology: A literature review. *Electronic Markets*, 32(4), 2245–2275. <https://doi.org/10.1007/s12525-022-00591-7>
- Li, W., Xu, S., Zheng, X., & Sun, R. (2024). Bridging the Knowledge Gap in Artificial Intelligence: The Roles of Social Media Exposure and Information Elaboration. *Science Communication*, 46(4), 399–430. <https://doi.org/10.1177/10755470241232352>
- Lin, C. H., Yu, C. C., Shih, P. K., & Wu, L. Y. (2021). STEM based Artificial Intelligence Learning in General Education for Non-Engineering Undergraduate Students. *Educational Technology & Society*, 24(3), 224–237.
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the Conference on Human Computer Interaction (HCI)*.

- Luttrell, R., Wallace, A., McCollough, C., & Lee, J. (2020). The Digital Divide: Addressing Artificial Intelligence in Communication Education. *Journalism & Mass Communication Educator*, 75(4), 470–482. <https://doi.org/10.1177/1077695820925286>
- Lutz, C. (2024). Social Inequalities and Artificial Intelligence: How Digital Inequality Scholarship Enhances Our Understanding. In D. Brzeziński, K. Filipek, K. Piwovar, & M. Winiarska-Brodowska (Eds.), *Algorithms, Artificial Intelligence and Beyond: Theorising Society and Culture of the 21st Century* (pp. 193–210). Routledge.
- MacInnis, D. J. (2011). A Framework for Conceptual Contributions in Marketing. *Journal of Marketing*, 75(4), 136–154. <https://doi.org/10.1509/jmkg.75.4.136>
- Malodia, S., Kaur, P., Ractham, P., Sakashita, M., & Dhir, A. (2022). Why do people avoid and postpone the use of voice assistants for transactional purposes? A perspective from decision avoidance theory. *Journal of Business Research*, 146, 605–618. <https://doi.org/10.1016/j.jbusres.2022.03.045>
- Mannuru, N. R., Shahriar, S., Teel, Z. A., Wang, T., Lund, B. D., Tijani, S., Pohboon, C. O., Agbaji, D., Alhassan, J., Galley, J., Kousari, R., Ogbadu-Oladapo, L., Saurav, S. K., Srivastava, A., Tummuru, S. P., Uppala, S., & Vaidya, P. (2023). Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. *Information Development*, Article 02666669231200628. Advance online publication. <https://doi.org/10.1177/02666669231200628>
- Mansell, R. (2017). The mediation of hope: Communication technologies and inequality in perspective. *International Journal of Communication*, 11, 4285–4304. <http://eprints.lse.ac.uk/81939/>
- Mayring, P., & Fenzl, T. (2014). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds.), *Handbuch. Handbuch Methoden der empirischen Sozialforschung* (pp. 543–556). Springer VS. https://doi.org/10.1007/978-3-531-18939-0_38
- McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., & Zolas, N. (2023). AI Adoption in America: Who, What, and Where. *National Bureau of Economic Research*. Advance online publication. <https://doi.org/10.3386/w31788>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. <https://pubmed.ncbi.nlm.nih.gov/23092060/>

- Mikalef, P [Patrick], Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and ‘the dark side’ of AI. *European Journal of Information Systems*, 31(3), 257–268. <https://doi.org/10.1080/0960085X.2022.2026621>
- Mikalef, P [Patrick], & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3).
- Mirbabaie, M., Brendel, A. B [Alfred B.], & Hofeditz, L. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1), 726–753. <https://doi.org/10.17705/1CAIS.05034>
- Mou, Y., & Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior*, 72, 432–440. <https://doi.org/10.1016/j.chb.2017.02.067>
- Nascimento, A., & Meirelles, F. (2022). An Artificial Intelligence Adoption Intention Model (AI2M) inspired by UTAUT. *ISLA 2022 Proceedings*. <https://aisel.aisnet.org/isla2022/21>
- Neumann, O., Guirguis, K., & Steiner, R. (2024). Exploring artificial intelligence adoption in public organizations: a comparative case study. *Public Management Review*, 26(1), 114–141. <https://doi.org/10.1080/14719037.2022.2048685>
- P. Norris, *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*, New York: Cambridge University Press, 2001.
- Oeldorf-Hirsch, A., & Neubaum, G. (2023). What do we know about algorithmic literacy? The status quo and a research agenda for a growing field. *New Media & Society*, Article 14614448231182662. Advance online publication. <https://doi.org/10.1177/14614448231182662>
- Oomen, T., Gonçalves, J., & Mols, A. (2023). Rethinking Artificial Intelligence: Algorithmic Bias and Ethical Issues| Rage Against the Artificial Intelligence? Understanding Contextuality of Algorithm Aversion and Appreciation. *International Journal of Communication*, 18(0), 25. <https://ijoc.org/index.php/ijoc/article/view/20809/4454>
- Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organization studies*, 28(9), 1435-1448.
- Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotko, M., Kacorri, H., Karwowski, W.,

- Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., . . . Xu, W. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human-Computer Interaction*, 39(3), 391–437. <https://doi.org/10.1080/10447318.2022.2153320>
- Pandl, K. D., Teigeler, H., Lins, S., Thiebes, S., & Sunyaev, A. (2021). Drivers and Inhibitors for Organizations' Intention to Adopt Artificial Intelligence as a Service. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 1769–1778).
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Park, S., & Humphry, J. (2019). Exclusion by design: intersections of social, digital and data exclusion. *Information, Communication & Society*, 22(7), 934–953. <https://doi.org/10.1080/1369118X.2019.1606266>
- Radhakrishnan, J., Gupta, S [Sumeet], & Prashar, S. (2022). Understanding Organizations' Artificial Intelligence Journey: A Qualitative Approach. *Pacific Asia Journal of the Association for Information Systems*, 14(6), 43–77. <https://doi.org/10.17705/1pais.14602>
- Ragnedda, M. (2017). *The third digital divide: A Weberian approach to digital inequalities*. Routledge advances in sociology: Vol. 204. Routledge Taylor & Francis Group.
- Rajagopal, A., & Vedamanickam, N. (2019). New Approach to Human AI Interaction to Address Digital Divide& AI divide: Creating an Interactive AIplatform to Connect Teachers & Students. In *Computer science & engineering: volume 2, Proceedings of 2019 Third IEEE International Conference on Electrical, Computer & Communication Technologies* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICECCT.2019.8869174>
- Riggins, F., & Dewan, S. (2005). The Digital Divide: Current and Future Research Directions. *Journal of the Association for Information Systems*, 6(12), 298–337. <https://doi.org/10.17705/1jais.00074>
- Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems*, 47(4), 425–436. <https://doi.org/10.1109/THMS.2017.2648849>
- Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M., & Stern, M. J. (2015). Digital inequalities and why they matter. *Information, Communication & Society*, 18(5), 569–582. <https://doi.org/10.1080/1369118X.2015.1012532>

- Sadek, M., Calvo, R. A., & Mougenot, C. (2024). Closing the Socio–Technical Gap in AI: The Need for Measuring Practitioners’ Attitudes and Perceptions. *IEEE Technology and Society Magazine*, 43(2), 88–91. <https://doi.org/10.1109/MTS.2024.3392280>
- Schaefer, C., Lemmer, K., Samy Kret, K., Ylinen, M., Mikalef, P [P.], & Niehaves, B. (2021). Truth or Dare? – How can we Influence the Adoption of Artificial Intelligence in Municipalities? In *Proceedings of the 54th Hawaii International Conference on System Sciences*. <https://scholarspace.manoa.hawaii.edu/handle/10125/72112>
- Scheerder, A., van Deursen, A., & van Dijk, J. (2017). Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. *Telematics and Informatics*, 34(8), 1607–1624. <https://doi.org/10.1016/j.tele.2017.07.007>
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2).
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109–124. <https://doi.org/10.17705/1thci.00131>
- Stolz, K., Hammerschmidt, T., & Posegga, O. (2024). How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers toward Sustainable Consumption: Exploring Nudges for Green Decision-Making. *Hawaii International Conference on System Sciences 2024 (HICSS-57)*. https://aisel.aisnet.org/hicss-57/in/social_shopping/3
- Stubbs, A. (2017). *Automation, Artificial Intelligence, and the God/Useless Divide*. <https://philpapers.org/rec/STUAAI>
- Templier, M., & Paré, G. (2018). Transparency in literature reviews: an assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, 27(5), 503–550. <https://ideas.repec.org/a/taf/tjisxx/v27y2018i5p503-550.html>
- Tornatzky, L. G., & Fleischer, M. (1990). *The Processes of Technological Innovation*. Lexington Books.
- Trajkovski, G. (2024). Bridging the public administration-AI divide: A skills perspective. *Public Administration and Development*, Article pad.2061. Advance online publication. <https://doi.org/10.1002/pad.2061>

- Tseng, C.-Y., & Ting, P.-H. (2013). Patent analysis for technology development of artificial intelligence: A country-level comparative study. *Innovation*, 15(4), 463–475.
<https://doi.org/10.5172/impp.2013.15.4.463>
- Uba, C., Lewandowski, T., & Böhmman, T. (2023). The AI-based Transformation of Organizations: The 3D-Model for Guiding Enterprise-wide AI Change. In T. Bui (Ed.), *Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 56th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2023.743>
- United Nations. (2020). *The Sustainable Development Report 2020*.
<https://sdgs.un.org/sites/default/files/2020-09/The-Sustainable-Development-Goals-Report-2020.pdf>
- Uren, V., & Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management*, 68, 102588.
<https://doi.org/10.1016/j.ijinfomgt.2022.102588>
- van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2022). Human-centred artificial intelligence: a contextual morality perspective. *Behaviour and Information Technology*, 41(3), 502–518.
<https://ideas.repec.org/a/taf/tbitxx/v41y2022i3p502-518.html>
- van Deursen, A., & Helsper, E. J. (Eds.). (2015). *Communication and Information Technologies Annual*. Emerald Group Publishing Limited.
- van Deursen, A., & van Dijk, J. (2011). Internet skills and the digital divide. *New Media & Society*, 13(6), 893–911. <https://doi.org/10.1177/1461444810386774>
- van Dijk, J. (2004). Divides in succession: Possession, skills, and use of new media for societal participation. *Media Access: Social and Psychological Dimensions of New Technology Use*, 233–254.
- van Dijk, J. (2008). One Europe, digitally divided. In A. Chadwick & P. N. Howard (Eds.), *Routledge handbook of internet politics* (0th ed., pp. 304–320). Routledge.
<https://doi.org/10.4324/9780203962541-28>
- Vassilakopoulou, P., & Hustad, E. (2023). Bridging Digital Divides: A Literature Review and Research Agenda for Information Systems Research. *Information Systems Frontiers*, 25(3), 955–969.
<https://doi.org/10.1007/s10796-020-10096-3>

- Venkatesh, V. (2022). Adoption and use of AI tools: a research agenda grounded in UTAUT. *Annals of Operations Research*, 308(1), 641–652. https://ideas.repec.org/a/spr/annopr/v308y2022i1d10.1007_s10479-020-03918-9.html
- Wagner, B. (2019). Ethics As An Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping? In E. Bayamlioglu (Ed.), *Being Profiled: 10 Years of Profiling the European Citizen* (1st ed., pp. 84–89). Amsterdam University Press. <https://doi.org/10.1515/9789048550180-016>
- Wamba, F. S., Queiroz, M. M., Pappas, I. O., & Sullivan, Y. (2024). Artificial Intelligence Capability and Firm Performance: A Sustainable Development Perspective by the Mediating Role of Data-Driven Culture. *Information Systems Frontiers*, 1–15. <https://doi.org/10.1007/s10796-023-10460-z>
- Wang, C.-H., & Wu, C.-L. (2022). Bridging the digital divide: the smart TV as a platform for digital literacy among the elderly. *Behaviour and Information Technology*, 41(12), 2546–2559. <https://ideas.repec.org/a/taf/tbitxx/v41y2022i12p2546-2559.html>
- Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & H de Vreese, C. (2024). The artificial intelligence divide: Who is the most vulnerable? *New Media & Society*, Article 14614448241232345. Advance online publication. <https://doi.org/10.1177/14614448241232345>
- Wang, J., Molina, M. D., & Sundar, S. S. (2020). When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior*, 107, 106278. <https://doi.org/10.1016/j.chb.2020.106278>
- Wang, Q., Baham, C., & Delen, D. (2024). The Impact of AI-Generated Misinformation Features on Human Believability: A Qualitative Approach. *AMCIS 2024 Proceedings*. https://aisel.aisnet.org/amcis2024/ai_aa/ai_aa/26
- Wang, S., Xiao, Y., & Liang, Z. (2024). Exploring cross-national divide in government adoption of artificial intelligence: Insights from explainable artificial intelligence techniques. *Telematics and Informatics*, 90, 102134. <https://doi.org/10.1016/j.tele.2024.102134>
- Wei, K.-K., Teo, H.-H., Chan, H. C., & Tan, B. C. Y. (2011). Conceptualizing and Testing a Social Cognitive Model of the Digital Divide. *Information Systems Research*, 22(1), 170–187. <https://doi.org/10.1287/isre.1090.0273>
- Weinert, L., Müller, J., Svensson, L., & Heinze, O. (2022). Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence

- Technologies in 40 German Hospitals: Descriptive Analysis. *JMIR Medical Informatics*, 10(6), e34678. <https://doi.org/10.2196/34678>
- Westerstrand, S. (2023). Ethics in the intersection of AI and democracy: The AIDEM Framework. *ECIS 2023 Research Papers*. https://aisel.aisnet.org/ecis2023_rp/321
- Winter, S., Berente, N., Howison, J., & Butler, B. (2014). Beyond the organizational ‘container’: Conceptualizing 21st century sociotechnical work. *Information and Organization*, 24(4), 250–269. <https://doi.org/10.1016/j.infoandorg.2014.10.003>
- Wu, Y. (2022). An Overview Analysis of AI Divide: Applications and Prospects of AI Divide in China’s Society. *2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022)*, 3–9. <https://doi.org/10.2991/assehr.k.220704.002>
- Xing, F., Peng, G., Zhang, B., Li, S., & Liang, X. (2021). Socio-technical barriers affecting large-scale deployment of AI-enabled wearable medical devices among the ageing population in China. *Technological Forecasting and Social Change*, 166, 120609. <https://doi.org/10.1016/j.techfore.2021.120609>
- Yang, J., Blount, Y., & Alireza Amrollahi (2021). Adoption of AI in the Auditing Practice: A Case study of a Big Four Accounting Firm. *ACIS 2021 Proceedings*. <https://aisel.aisnet.org/acis2021/26>
- Young, E., Wajcman, J., & Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, 38(3), 391–414. <https://doi.org/10.1111/ntwe.12278>
- Yuan, Y., Sun, Y., & Chen, H. (2024). Does artificial intelligence affect firms’ inner wage gap? *Applied Economics*, 1–7. <https://doi.org/10.1080/00036846.2024.2324090>
- Zhang, J. J. Y., Følstad, A., & Bjørkli, C. A. (2023). Organizational Factors Affecting Successful Implementation of Chatbots for Customer Service. *Journal of Internet Commerce*, 22(1), 122–156. <https://doi.org/10.1080/15332861.2021.1966723>
- Zhao, C., Li, Y., Liu, Z., & Ma, X. (2024). Artificial intelligence and carbon emissions inequality: Evidence from industrial robot application. *Journal of Cleaner Production*, 438, 140817. <https://doi.org/10.1016/j.jclepro.2024.140817>

Appendix I: Digital transformation affecting divide research.

<i>Phases</i>	<i>Digital Transformation</i>	<i>Changes in Digital Divide Research</i>
<i>Analytics 1.0 – Business Intelligence</i>	- collection, aggregation, and analysis of manufacturing data, sales and customer interactions - especially available for larger companies, due to heavy investments - new skills: data preparation rather than data analysis - historical focus, no predictions - few people were talking about competitive advantage from analytics	- the digital divide as “the separation between those who have access to digital information and communication technology (ICT) and those who do not” was investigated in the context of business management research (Riggins & Dewan, 2005, p. 298) - the impact of the divide on society and organizations depends on financial spending and is rarely discussed (Riggins & Dewan, 2005)
<i>Analytics 2.0 – Big Data</i>	- since about the mid-2000s - inclusion of internal as well as external data sources (e.g., internet, audio, video recordings, sensors) - first-mover advantage led to new opportunities and technologies, which were created - analysis of data by using machine learning methods	- distinguishing the digital divide into the access divide, capability divide, and the outcome divide resulting in a change from a dichotomous view (access yes or no) to a more qualitative view (digital inclusion vs. exclusion) (van Deursen & van Dijk, 2011; Wei et al., 2011)
<i>Analytics 3.0 – Data economy analytics</i>	- customized products and services through search algorithms on websites and targeted advertising - also, companies in traditional industries make use of analytics and big data - longstanding companies are in need to transform their business and cultures - analytical activities are getting “industrialized”	- algorithmic divide (Klumpp, 2017) - Big Data allows novel forms of AI applications and platform services (Klumpp, 2017; Rajagopal & Vedamanickam, 2019) - the divide can be characterized by a gap between humans that can benefit from the potential of AI for human progress and critically understanding its threats and those who cannot (Rajagopal & Vedamanickam, 2019) - AI is crucial in knowledge economies to achieve learning experience and efficiency due to the use of applications and systems (Tseng & Ting, 2013)
<i>Analytics 4.0 – Artificial Intelligence</i>	- IT knowledge often goes beyond traditional analytical skills; however frequently machine learning as predictive analytics is at the core of AI - models help companies to develop new markets, products and services and help to improve quality to adapt to changing market environments - a mix of data analytic skills, programming and business development is important	- the greater level of interactivity and connectivity between AI-based technologies and humans is changing the divide (e.g., through interactive platforms or robotics (Chang et al., 2022) - the autonomy of AI transforms the digital divide into an AI divide (Carter et al., 2020; Wu, 2022)

Table I.1: Four phases of AI development, based on Davenport (2018).

Appendix II: Literature screening.

II.1. Format screening.

Database	Non-English	Non-scientific	Incomplete	Literature review	Total
Definition	Articles not completely written in English.	Gray literature, book chapters, book reviews, dissertations, tutorials, bachelor's or master's theses, workshops, posters, letter of editors, special issues, symposiums, talks, magazine.	Short papers or research-in-progress papers, extended abstracts.	Standalone literature reviews (no articles combining a literature review with another methodology).	-
ACM	0	37	4	8	49
AIS	0	488	21	14	523
BSC	19	638	41	14	712
Scholar	1	59	5	11	76
CSDL	0	121	22	17	160
SciDir	24	448	0	11	483
WoS	6	437	6	312	761
Total	50	2,225	99	387	2,764

Table II.1 Number of excluded articles through format screening.

II.2 Relevance screening.

Database	No AI focus	No divide focus	No SST focus	Total
Definition	Technologies mentioned in the articles must carry out tasks that typically require human cognition.	Articles must provide dedicated contributions regarding factors, drivers, barriers, or circumstances that lead to the inclusion (reducing the gap/divide) or exclusion (fostering the gap/divide) between AI-enabled individuals, organizations, or society from participation and non-enabled ones.	Articles must consider human-AI interaction within a sociotechnical system following the conceptualization of Leonardi (2012) – meaning that articles should not solely focus on adopting certain working processes toward specific AI-related tasks by considering ex-ante and ex-post process besides the interaction itself.	-
ACM	14	98	16	128
AIS	14	180	201	395
BSC	109	258	69	436
CSDL	4	260	524	788
Scholar	0	7	2	9
SciDir	705	1,709	679	3,093
WoS	234	591	69	894
Total	1,080	3,103	1,560	5,743

Table II.2 Number of excluded articles through relevance screening.

Appendix III: Final sample.

Database	Keyword Search	Backward & Forward Search	Total
ACM	9	4	13
AIS	2	0	2
BSC	7	1	8
CSDL	0	0	0
Scholar	3	0	3
SciDir	8	2	10
WoS	17	0	17

Others	-	6	6
Total	46	13	59

Table III.1 Distribution of final sample among databases including backward and forward searches.

Autor(s)	Year	Title	Outlet	Source *
Al-Fraihat et al.	2023	Why do Consumers Adopt Smart Voice Assistants for Shopping Purposes? A Perspective From Complexity Theory	Intelligent Systems with Applications	FS 3 rd
Al Shamsi et al.	2022	Understanding Key Drivers Affecting Students' use of Artificial Intelligence-Based Voice Assistants.	Education and Information Technologies	FS 2 nd
Alsheibani et al.	2020	Artificial Intelligence Beyond the Hype: Exploring the Organisation Adoption Factors	Americas Conference on Information Systems (AMCIS)	KS
Alsheibani et al.	2019	Factors Inhibiting the Adoption of Artificial Intelligence at Organizational-level: A Preliminary Investigation	Australasian Conference on Information Systems (ACIS)	KS
Aydin et al.	2023	Does Industry 5.0 Reproduce Gender (In)equalities at Organisations? Understanding the Interaction of Human Resources and Software Development Teams in Supplying Human Capitals	Information Systems Frontiers	KS
Bérubé et al.	2021	Barriers to the Implementation of AI in Organizations: Findings From a Delphi Study	Hawaii International Conference on System Sciences (HICSS)	KS
Celik	2023	Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption	Telematics and Informatics	KS
Daepf and Counts	2024	The Emerging AI Divide in the United States	arXiv preprint	KS
Diel et al.	2021	Alexa – Welcome to the Family! IT Identity's Mediating Role on Social Presence and Deep Use of In-Home Voice Assistants	International Conference on Information Systems (ICIS)	KS
García de Blanes Sebastián et al.	2022	Application and Extension of the UTAUT2 Model for Determining Behavioral Intention Factors in use of the Artificial Intelligence Virtual Assistants	Frontiers in Psychology	FS 3 rd
Frick et al.	2021	Maneuvering Through the Stormy Seas of Digital Transformation: the Impact of Empowering Leadership on the AI Readiness of Enterprises	Journal of Decision Systems	FS 1 st
Gao	2024	Shadow of Technology: Robotics Development and Regional Economic Inequality	Journal of Knowledge Economy	KS
Gignac and Szodorai	2024	Defining Intelligence: Bridging the gap Between Human and Artificial Perspectives	Intelligence	KS
Goldenthal et al.	2021	Not All AI are Equal: Exploring the Accessibility of AI-Mediated Communication Technology	Computers in Human Behavior	KS
Grundspenki	2007	Agent Based Approach for Organization and Personal Knowledge Modelling: Knowledge Management Perspective	Journal of Intelligent Manufacturing	KS
Hansen et al.	2022	Examining AI Adoption Through a Maturity Model: A Case Study on Norwegian Organizations	MENA Conference on Information Systems (MENACIS)	FS 1 st
Hradecky et al.	2022	Organizational Readiness to Adopt Artificial Intelligence in the Exhibition Sector in Western Europe	International Journal of Information Management	KS
Holtel	2016	Artificial Intelligence Creates a Wicked Problem for the Enterprise	Procedia Computer Science	KS
Hsiao and Han	2023	The Impact of Data Analytics and Artificial Intelligence on the Future Profession: Perspectives From Accounting Students	Journal of Theoretical Accounting Research	KS
Iuga and Socol	2024	Government Artificial Intelligence Readiness and Brain Drain: Influencing Factors and Spatial Effects in the European Union Member States	Journal of Business Economics & Management	KS

Jain et al.	2022	Adoption of AI-Enabled Tools in Social Development Organizations in India: An Extension of UTAUT Model	Frontiers in Psychology	FS 1 st
Jeon et al.	2024	Speed of Catch-Up and Convergence of the Artificial Intelligence Divide: AI Investment, Robotic, Start-Ups, and Patents	Journal of Global Information Technology Management	KS
Jöhnk et al.	2021	Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors	Business & Information Systems Engineering	FS 1 st
Josifidis and Supic	2024	Artificial Intelligence and the Productivity-Pay Gap in the USA: Industrial Insights and the Revival of Heterodox Ideas	Journal of Economic Issues	KS
Kamoonpuri and Sengar	2023	Hi, May AI Help you? An Analysis of the Barriers Impeding the Implementation and use of Artificial Intelligence-Enabled Virtual Assistants in Retail	Journal of Retailing and Consumer Services	FS 1 st
Klumpp and Zijm	2019	Logistics Innovation and Social Sustainability: How to Prevent an Artificial Divide in Human-Computer Interaction.	Journal of Business Logistics	KS
Kutz et al.	2022	Implementation of AI Technologies in Manufacturing – Success Factors and Challenges	The Human Side of Service Engineering	FS 1 st
Laut et al.	2021	Integration of Artificial Intelligence in the Organizational Adoption – A Configurational Perspective	International Conference on Information Systems (ICIS)	KS
Lainjo	2023	The Global Social Dynamics and Inequalities of Artificial Intelligence	International Journal of Innovation Scientific Research and Review	KS
W. Li et al.	2024	Bridging the Knowledge Gap in Artificial Intelligence: The Roles of Social Media Exposure and Information Elaboration	Science Communication	KS
Malodia et al.	2022	Why do People Avoid and Postpone the use of Voice Assistants for Transactional Purposes? A Perspective from Decision Avoidance Theory	Journal of Business Research	KS
Mannuru et al.	2024	Artificial Intelligence in Developing Countries: The Impact of Generative Artificial Intelligence (AI) Technologies for Development	Information Development	KS
Mansell	2017	The Mediation of Hope: Communication Technologies and Inequality in Perspective	International Journal of Communication	KS
McElheran et al.	2024	AI Adoption in America: Who, What, and Where	Journal of Economics & Management Strategy	KS
Mirbabaie et al.	2022	Why do People Avoid and Postpone the use of Voice Assistants for Transactional Purposes? A Perspective From Decision Avoidance Theory	Journal of Business Research	FS 2 nd
Mikalef et al.	2022	Enabling AI Capabilities in Government Agencies: A study of Determinants for European Municipalities	Government Information Quarterly	KS
Mou and Xu	2017	The Media Inequality: Comparing the Initial Human-Human and Human-AI Social Interactions	Computers in Human Behavior	KS
Nascimento and Meirelles	2022	An Artificial Intelligence Adoption Intention Model (AI2M) Inspired by UTAUT	Information Systems in Latin America (ISLA)	KS
Neumann et al.	2022	Exploring Artificial Intelligence Adoption in Public Organizations: a Comparative Case Study	Public Management Review	KS
Oeldorf-Hirsch and Neubaum	2023	What do we Know About Algorithmic Literacy? The Status quo and a Research Agenda for a Growing Field	New Media & Society	KS
Oomen et al.	2024	Rage Against the Artificial Intelligence? Understanding Contextuality of Algorithm Aversion and Appreciation	International Journal of Communication	KS
Pandl et al.	2021	Drivers and Inhibitors for Organizations' Intention to Adopt Artificial Intelligence as a Service	Hawaii International Conference on System Sciences (HICSS)	KS
Park and Humphry	2019	Exclusion by Design: Intersections of Social, Digital and Data Exclusion	Information, Communication & Society	KS

Radhakrishnan et al.	2022	Understanding Organizations' Artificial Intelligence Journey: A Qualitative Approach	Pacific Asia Journal of Association for Information Systems	FS 1 st
Sadek et al.	2023	Closing the Sociotechnical Gap in AI: The Need for Measuring Practitioners' Attitudes and Perceptions	IEEE Technology and Society Magazin	KS
Schaefer et al.	2021	Truth or Dare? – How can we Influence the Adoption of Artificial Intelligence in Municipalities?	Hawaii International Conference on System Sciences (HICSS)	KS
Stubbs	2017	Automation, Artificial Intelligence, and the God/Useless Divide	Perspectives on Global Development & Technology	KS
Trajkovski	2024	Bridging the Public Administration-AI divide: A Skills Perspective	Public Administration and Development	KS
Tseng and Ting	2013	Patent Analysis for Technology Development of Artificial Intelligence: A Country-Level Comparative Study	Innovation: Management, Policy & Practice	KS
Uba et al.	2023	The AI-based Transformation of Organizations: The 3D-Model for Guiding Enterprise-wide AI Change	Hawaii International Conference on System Sciences (HICSS)	FS 1 st
Uren and Edwards	2023	Technology Readiness and the Organizational Journey Towards AI Adoption: An Empirical Study	International Journal of Information Management	KS
S. Wang et al.	2024a	Exploring Cross-National Divide in Government Adoption of Artificial Intelligence: Insights From Explainable Artificial Intelligence Techniques	Telematics and Informatics	KS
C. Wang et al.	2024b	The Artificial Intelligence Divide(Brock & Wangenheim, 2019): Who is the Most Vulnerable?	New Media & Society	KS
Weinert et al.	2022	Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence Technologies in 40 German Hospitals: Descriptive Analysis	MIR Medical Informatics	KS
Xing et al.	2021	Sociotechnical Barriers Affecting Large-Scale Deployment of AI-Enabled Wearable Medical Devices Among the Ageing Population in China	Technological Forecasting and Social Change	KS
Yang et al.	2021	Adoption of AI in the Auditing Practice: A Case study of a Big Four Accounting Firm	Australasian Conference on Information Systems (ACIS)	KS
Young et al.	2023	Mind the Gender gap: Inequalities in the Emergent Professions of Artificial Intelligence (AI) and Data Science	New Technology Work and Employment	KS
Yuan et al.	2024	Does Artificial Intelligence Affect Firms' Inner Wage gap?	Applied Economics	KS
Zhang et al.	2023	Organizational Factors Affecting Successful Implementation of Chatbots for Customer Service	Journal of Internet Commerce	FS 2 nd
Zhao et al.	2024	Artificial Intelligence and Carbon Emissions Inequality: Evidence From Industrial Robot Application	Journal of Cleaner Production	KS

*KS = Keyword-Search, BS = Backward-Search, FS = Forward-Search

Table III.2 Overview of the final sample.

Appendix IV: Snapshot of coding.

Part of the Sociotechnical System	Part of the Conceptualization	Code	Coding Example
Social subsystem (Societal/geographical perspective)	Human-related objects	Societal culture	“Socio-cultural challenges represent the issues that are rooted in society and culture [...]that must be considered while designing, developing, and deploying virtual assistants” (Kamoonpuri & Sengar, 2023, 2,4)
		Public trends	“...organizations has recently incorporated new technology to follow the stated trend...” (Hradecky et al., 2022, p. 8)

	Non-human objects	Infrastructure	“Low level of IT infrastructure (limited bandwidth, inadequate technology and systems, poor telecommunications structure, lack of computing equipment, insufficient resources)” (Bérubé et al., 2021, p. 6704)
		Societal governance	“[E]thics and data privacy, for example, are primarily social issues which require organizations to establish data governance to guide technological solutions.” (Uren & Edwards, 2023, p. 9)
		Public funding	“However, for a developing country like China with 176 million of elderly people, public funding will never be sufficient to meet the huge public demand...” (Xing et al., 2021, p. 21)
Social subsystem (Industrial/market perspective)	Human-related objects	Innovative potential	“Clinical barriers [...] innovative and high-quality [wearable medical devices] that have obtained clinical approval...” (Xing et al., 2021, p. 18)
		Customer properties	“Our findings further suggest that existing nudging strategies are not sufficiently effective in addressing consumers’ decision avoidance, specifically customer inertia. Hence, marketers must devise innovative nudging strategies to reduce procrastination behaviour and overcome customer inertia.” (Malodia et al., 2022, p. 615)
	Non-human objects	Industry characteristics	“Reflecting the structure of the broader events industry in general and business events in particular, the exhibition industry is fragmented and made up of many different sectors, organizations and suppliers [...] The evolution of the exhibition industry is closely connected with technological development. However, the experience or level of readiness and intention that exhibition suppliers have to implement emerging technologies such as AI remains under-researched. (Hradecky et al., 2022, p. 3)
		Industrial governance	“A general skepticism toward AI is noticeable due to the highly restricted industry that they operate in (Bank2).” (Hansen et al., 2022, p. 9)
		Proof of concept	“...marketers can proactively create social proof through advertisements of how VAs [Virtual Assistants] can be used in the purchase cycle. There is a need to either integrate them into the purchase journey or establish them as a breakthrough, atypical medium of purchase.” (Kamoonpuri & Sengar, 2023, p. 11)
Social subsystem (Institutional/organizational perspective)	Human-related objects	Strategic leadership	“At the leadership level, a lack of understanding of AI on the part of management creates barriers to AI implementation. Our respondents differentiated between a lack of understanding of the technical aspects of AI and of the business potential of AI as separate barriers, with a lack of understanding of the business potential of AI being the highest ranked barrier by our respondents.” (Bérubé et al., 2021, p. 6708)
		Corporate culture	“This outcome can be understood by the fact that there are likely other aspects that exert a stronger impact on ability of municipalities to develop an AI capability, such as a culture for innovativeness and the right mix of incentives and push from higher government, which likely render perceptions of value as less important.” (Mikalef et al., 2022, p. 8)

	Non-human objects	Institutional resources	“Two critical resources discussed by interviewees were financial resources and AI-human resources. Financial resources referred to the capital budget that the firm could spend on developing the AI enabled auditing tools.” (Yang et al., 2021, p. 7)
		Institutional governance	“Furthermore, there is a lack of frameworks to guide the ethical management of AI in profit and non-profit organizations.” (Mirbabaie et al., 2022, p. 747)
Social subsystem (Unit/team perspective)	Human-related objects	Social Influence	“As per social identity theory, AI collaborative tools are used by team members, and if some members of the team adopt the technology, there would be a form of social pressure on others to adopt the technology, or else the team members might feel left out” (Jain et al., 2022, p. 10)
	Non-human objects	Financial power	“...but there are lower levels of affordability among this group than in the wider population” (Park & Humphry, 2019, p. 936)
Sociomaterial practices (Social entity: human user)	Skill-related characteristics	AI skills	“Lack of AI skills to implement and manage AI were identified as the most common barriers (33.8%)” (Alsheibani et al., 2019, p. 6)
	Sociodemographic characteristics	Age	“Encouraging customer adoption of [smart voice assistants] SVAs in Jordan should focus on motivating younger females with lower educational levels and those with low income to engage in adoption” (Al-Fraihat et al., 2023, p. 10)
	Personality-related characteristics	Habits	“However, a profound understanding of IT identity-induced deep use of VAs is still missing, despite VAs' potential to disrupt whole generations' working habits and everyday life.” (Diel et al., 2021, p. 2)
	Perception-related characteristics	Self-efficacy	“When dealing with new technology adoption (e.g., AI), people face challenges, and people with higher self-efficacy would persist and engage more than those with lower self-efficacy.” (Nascimento & Meirelles, 2022, p. 8)
Sociomaterial practices (Material entity: AI)	Artifact perspective	System quality	“Our initial qualitative inquiry also helped us understand the potential moderating role of technology comfort in the relationship between cognitive biases and the decision to use [virtual assistants] VAs for transactions (i.e. the rejection of VAs for transactions or the intention to adopt VAs for transactions)” (Malodia et al., 2022, p. 610)
		System transparency	“Another factor that increases the complexity of AIaaS [as a service] is its opaque data processing, which makes it challenging to understand...” (Pandl et al., 2021, p. 1775)
		AI governance	“AI-based systems are at risk for biased learning and unethical outcomes” (Jöhnk et al., 2021, p. 11)

Table V.1 Snapshot of coding.

Paper IX

Hammerschmidt, T.; Passlack, N. & Posegga, O. (working paper, unpublished). *How Organizations Ethically Respond to Decision Dilemmas in AI Deployment*. [previously 2nd revision round at JSIS]

How Organizations Ethically Respond to Decision Dilemmas in AI Deployment

Working Paper (unpublished)

Abstract

Within organizations there is a growing deployment of artificial intelligence (AI) technologies – either as a strategic choice planned by the management to stay competitive by increasing working efficiency and business value (top-down AI deployment) with AI or unsupervised by having employees using free-to use AI tools to optimize their daily work (bottom-up AI deployment). These ways of AI deployment change how organizational stakeholders, such as employees and the management board, develop different strategies to deal with the implications of AI collaboration. Particularly potential negative implications challenges organizations to harmonize sometimes conflicting stakeholder choices when applying AI to their business strategies. Through qualitative interviews, this study highlights the importance of ethical considerations when weighting between different organizational strategies for AI deployment to deal with potential negative implications. It outlines how organizations respond to decision dilemmas that arise when different stakeholder choices on how to collaborate with AI ethically are in conflict and, thereby, providing ideas on how to extend the concept of ambidexterity for ethical AI deployment strategies.

Keywords: *Ambidexterity; Ethics; Human-AI interaction; AI Deployment, Future of Work, Organizational Theory and Change*

1. Introduction

Artificial intelligence (AI) technologies—ranging from generative tools and predictive analytics to robotics—are increasingly used to optimize operational processes, collaboration working tasks, and automate or augment decision-making in organizations (Collins et al., 2021). These AI systems perform cognitive functions traditionally associated with human intelligence, such as reasoning, learning, and perceiving (Rai et al., 2019), thereby driving changes in organizational power dynamics, collaboration norms, workers' roles, and responsibility structures (Giermindl et al., 2022; O'Neill et al., 2022). Such changes impact how different

organizational stakeholders collaborate with AI technologies and deal with its implications, making the identification of suitable strategies for AI deployment a critical challenge (Chanias et al., 2019; Vial, 2019).

We understand AI deployment strategies as the deliberate and iterative processes of formulating, implementing, and adapting strategies to deploy AI to achieve an organization's long-term objectives, mitigating potential risks, and align it with its stakeholders values and norms (Ayoub & Payne, 2016; Bareis & Katzenbach, 2022). Each strategy entails specific strategic choices, including decisions about where, when, and how to deploy AI (Alsheibani et al., 2020; Brock & Von Wangenheim, 2019). This involves navigating complex *decision dilemmas* as weighing between several conflicting choices that carry the risk of unavoidable, unintended, and negative side effects for the organization or some of its stakeholders (Bonnefon et al., 2024; Kegel & Ghanem, 2024; Marabelli & Galliers, 2017; Shank et al., 2019). For instance, AI may “reduce the costs for a firm and thus bring performance benefits such as improved profitability, [but] can also demoralize the human resources as they would perceive themselves to be under the continuous threat of being replaced by machines” (Pan et al., 2024, p. 3). Organizations must, therefore, balance organizational stakeholder's (potentially conflicting) choices to reduce unintended negative implications for individual employees and the organization as a whole (Enholm et al., 2022; Morton et al., 2022; Wimelius et al., 2021).

Conflicting choices might arise due to the widespread use of easily accessible and free-to-use AI applications through employees in workplace context, whereby AI deployment is no longer solely guided by top-down strategies (Berente et al., 2019; Dwivedi et al., 2021, 2023). Instead, it often emerges bottom-up, driven by employees experimenting with AI in ways that are partially unsupervised or even invisible to organizations (Hammerschmidt et al., 2024). This decentralized adoption increases the complexity of finding appropriate AI deployment strategies by the management (top-down), which is further heightened by the dual strategic logic organizations must manage: *exploitation*—optimizing existing processes and structures—and *exploration*—pursuing innovation and business value (A. Chen et al., 2023; Wirtz, 2020). This balancing act, known as ambidextrous decision-making, becomes more complex in the context of AI, where outcomes of such strategic choices are difficult to predict and ethical implications are not always clear (Ling et al., 2022; Shahid et al., 2023). For instance, exploitation strategies may maximize production efficiency by reducing costs (e.g., AI's increasing autonomy replacing human workers) but risk alienating workers' job security (Bankins &

Formosa, 2023; Marabelli & Galliers, 2017). Conversely, exploration strategies may promote innovation but lead to regulatory uncertainty and role conflicts of employees (Hammerschmidt et al., 2025; Mirbabaie, Brünker, et al., 2022). Consequently, dealing with decision dilemmas within AI deployment involves navigating ethical considerations while striking a balance between exploration and exploitation strategies, which challenges previous assumptions about ambidexterity (Cao et al., 2024; Van De Wetering et al., 2022).

Despite a growing body of valuable research that integrates ethical considerations into the strategic management of human-AI interactions in organizations (e.g., Bareis & Katzenbach, 2022; Brendel et al., 2021; Heyder et al., 2023), the integration of ethical considerations in ambidexterity concepts has yet to be explored in detail (Hammerschmidt et al., 2024; Ling et al., 2022; Shahid et al., 2023). This is particularly the case when aiming to better understand the individual goals of organizational stakeholders forming different strategic choices on how to collaborate with AI and how to deal with its implications (Cao et al., 2024; Shahid et al., 2023; Van De Wetering et al., 2022). Our research questions (RQs) address this gap by asking: *What decision dilemmas do organizations face within their exploration and exploitation strategies when deploying AI (RQ1)?* And, *how do they respond to potential decision dilemmas within AI deployment ethically (RQ2)?*

To achieve our research goal, we conducted 27 interviews with both leaders and employees to explore different strategic choices organizations confront when deploying AI into their operations and workforce. The semi-structured interviews allowed us to identify different decision dilemmas of organizational stakeholders when deploying AI and how organizations respond to these dilemmas. Our explorative analysis focuses on experiences with AI deployment in organizations, interviewing leaders and employees from various disciplines and organizations. Thereby, we expect to contribute to Information Systems (IS) research in three ways: *First*, we follow the calls of Ling et al. (2022), Shahid et al. (2023) and Yang et al. (2024) by providing ideas on how to integrate ethical considerations into previous concepts of ambidexterity based on an in-depth examination of real-world AI deployment scenarios, in which organizations develop suitable strategies to integrate AI into their working processes ethically by balancing between empowering its stakeholders for responsible AI collaboration while preserving managerial control over its implications. *Second*, we follow the calls of Hafner et al. (2024), Hammerschmidt et al., (2025) and Mirbabaie et al. (2022) for more qualitative research that incorporates human stakeholders' virtue ethical perspectives to advance AI ethics literature in IS research by examining four key

decision dilemmas arising in AI deployment through the lens of our participants' inherent values and ethical perceptions. *Third*, we follow the calls of Alsheiabni et al. (2020), Mirbabaie et al. (2022), Pan et al. (2024), Vial (2019) and Zhang et al. (2022) by providing empirical insights on why and how different organizational stakeholders follow different strategic approaches when seeking to collaborate with AI and its implications in an ethical manner and by providing a practice-related discussion on how organizations may respond to such potential decision dilemmas.

In section 2, we outline the background that drives our empirical investigation. We then describe, in Section 3, the methods underlying our qualitative approach. Section 4 presents the results. This is followed by a discussion of the theoretical and practical implications of our results and the limitations in Section 5. Section 6 concludes with a summary.

2. *Theoretical Background: Why we need ethical considerations within ambidextrous decision-making*

This section introduces several concepts of ambidexterity and decision dilemmas in AI deployment strategies, with particular emphasis on the relevance for ethical considerations that arise at their intersection.

Concepts of Ambidexterity

In the realm of strategic decision-making of organizations, existing research commonly distinguishes between two fundamental types of organizational strategy: *exploitation*, which involves refining existing practices and processes to enhance efficiency and optimize current business performance, and *exploration*, which focus on the pursuit of novel opportunities and innovative approaches to create future business value (Y. Chen et al., 2022; Marabelli & Galliers, 2017). Striking a balance between these two strategies is crucial as too much focus on exploration can lead to costs for experimentation without reaping benefits, whereas too much focus on exploitation can hinder innovation (O'Reilly & Tushman, 2004). Balancing exploration and exploitation is typically referred to as ambidexterity, which was introduced first by March (1991) and extended by O'Reilly and Tushman (2004, 2011). Ambidexterity refers to an organization's ability to simultaneously pursue both exploration and exploitation, despite the inherent tensions and resource constraints involved (Mom et al., 2019). Achieving such a balance is particularly challenging in times of technological disruption and

transformation, where organizational resources are often limited and the demands for adaptability are heightened (Mom et al., 2019). Thus, we understand ambidextrous decision-making in this work as the organizational capability to make strategic choices that integrate or shift between exploration and exploitation in alignment with internal demands and resources.

There exist various approaches to achieve this balance in literature – linked to different concepts of ambidexterity. One prominent form is *structural ambidexterity*, wherein exploration and exploitation are compartmentalized into distinct organizational units (e.g., De Visser et al., 2010). For instance, research and development (R&D) departments may focus on exploration to identify novel business potentials, while human resource management (HRM) focuses on exploitation to refine and optimize internal processes and knowledge (Cao et al., 2024). This structural separation allows organizations to pursue both strategic logics simultaneously, without one undermining the other. An alternative approach is *sequential ambidexterity*, which involves temporal shifts between periods of exploration and periods of exploitation (e.g., Goossen et al., 2012). Rather than pursuing both strategies concurrently, organizations may alternate between them—for instance, initiating short phases of radical innovation in specific projects followed by longer periods of focused on improvements of daily working routines. This temporal sequencing enables firms to manage resource constraints while still engaging with both strategies.

Punctuated equilibrium, is another alternative concept, borrowed from evolutionary biology (Eldredge, 1972), that offers a contrasting perspective to ambidexterity (Gupta et al., 2006). Rather than emphasizing balance between exploration and exploitation, this concept suggests that organizations remain relatively stable for long periods until being interrupted by brief, radical transformations response to environmental shocks or internal crises (Romanelli & Tushman, 1994). Unlike sequential ambidexterity, which assumes a planned or iterative shift between exploration and exploitation, punctuated equilibrium underscores the reactive and often abrupt nature of organizational adaptation. This distinction highlights the fundamentally different temporal and structural dynamics at play: while ambidexterity relies on continuous or planned alternation between strategic modes, punctuated equilibrium reflects a more disruptive and externally triggered model of organizational change (Gupta et al., 2006, O'Reilly & Tushman, 2004).

In our qualitative analysis, we engage with the diverse concepts surrounding ambidexterity literature—namely structural ambidexterity, sequential ambidexterity, and the punctuated equilibrium idea—recognizing that each offers a distinct lens for understanding how organizations face and manage the competing strategic choices. While our analysis does not exhaustively cover the full breadth of the ambidexterity literature, it draws on the most relevant ones that undermine the theoretical implications of our discussion.

Ethical Decision Dilemmas Within AI Deployment

AI deployment can be understood as a deliberate and iterative process of integrating artificial intelligence (AI) technologies into organizational operations, with the aim of achieving long-term strategic objectives, optimizing performance, and aligning technological implementation with the values and expectations of internal and external stakeholders (Ayoub & Payne, 2016; Bareis & Katzenbach, 2022; Dwivedi et al., 2021). AI deployment strategies encompass a range of strategic decisions concerning where, when, and how AI technologies should be implemented (Alsheibani et al., 2020; Brock & von Wangenheim, 2019). These strategic decisions of AI deployment are closely tied to ambidexterity: AI is deployed not only to explore new sources of value and innovation (linked to *exploration*) but also to exploit existing capabilities to improve efficiency and productivity (linked to *exploitation*).

However, AI deployment often entails navigating complex decision dilemmas as organizations must weigh competing priorities—such as efficiency versus fairness or automation versus human oversight—which arise when AI strategies do not align with heterogeneous stakeholder goals (Bonnefon et al., 2024; Marabelli & Galliers, 2017; Shank et al., 2019). These dilemmas are characterized by situations in which any chosen course of action may produce adverse outcomes for certain organizational stakeholders, raising the risk of unintended negative consequences (Bonnefon et al., 2024; Kegel & Ghanem, 2024). For instance, a decision to implement AI for workforce optimization (exploitation) may increase organizational efficiency but simultaneously lead to job displacement, creating tensions between economic and social goals (Milanez, 2024). Conversely, focusing on exploration may lead to disruptive innovations that benefit the organization but destabilize industries or displace employees (Shankar, 2024). Thus, we argue that decision dilemmas in ambidextrous strategies are by

their nature ethical decision dilemmas because they require aligning strategic choices with ethical considerations.

Thus, AI deployment strategies extend beyond the dichotomy of balancing exploration and exploitation. It demands the integration of ethical considerations on required governance structures or the alignment of individual values and organizational norms during AI deployment initiatives (Jöhnk et al., 2021; Sousa et al., 2022); ethical considerations with respect to ambidextrous decision-making that have not been explored in detail to date (Hammerschmidt et al., 2024; Ling et al., 2022; Shahid et al., 2023). In this work, we understand ethical considerations as the evaluation of what is right or wrong, fair or unfair, and beneficial or harmful in the eyes of different organizational stakeholders, particularly when decisions have implications for individual organizational stakeholders and organizations at large (Bankins & Formosa, 2023; Floridi & Cowls, 2019; Hagendorff, 2020).

Decision dilemmas within AI deployment are further intensified by the tension between top-down and bottom-up implementation dynamics. Traditionally, ambidextrous decision-making has been regarded as a top-down managerial tasks, wherein leaders simultaneously address the competing demands of exploration and exploitation within their organizations (O'Reilly & Tushman, 2011; Shahid et al., 2023). However, in practice, AI deployment is increasingly driven from the bottom-up – initiated by employees who engage with these technologies in proactive, often unsupervised ways that precede formal strategies or policies (Dwivedi et al., 2023; Maedche et al., 2019). For instance, in exploration-oriented working contexts, individual employees may adopt AI tools to engage with disruptive innovations and secure their future employability, often through experimentation and skills development (Chan et al., 2019). In contrast, within exploitation context, employees may pursue alignment-oriented activities aimed at optimizing existing processes, enhancing their value in daily operations, or conserving personal resources such as time and effort (Chan et al., 2019). This decentralized, bottom-up engagement does not simply supplement top-down strategic initiatives—it can amplify decision dilemmas by introducing competing priorities, value conflicts, and unintended consequences that emerge from a misalignment between individual and organizational goals.

In our qualitative analysis, we investigate a range of decision dilemmas that emerge in real-world AI deployment strategies, outlining the relevance of ethical considerations that organizations must navigate when implementing AI technologies dynamic organizational contexts.

3. Method

To identify decision dilemmas practitioners face when making strategic choices, and to identify how organizations can respond to them ethically, we conducted semi-structured, qualitative interviews with experts who deal with AI deployment and use in organizations. The qualitative data helped us derive insights from the perspective of different expert's experiences and perceptions that are created through social processes of interacting with the technology and the organizational environment (Sarker et al., 2018). This is an interpretive form of qualitative research that mainly follows an inductive approach (Sarker et al., 2018).

Data collection. We recruited two groups for our interviews: leaders whose roles are driving AI deployment in organizations and aligning AI strategies with organizational resources to achieve strategic goals (e.g., Watson et al., 2021); and employees working with AI technologies to complete their daily tasks or increase productivity and are thus driving unsupervised AI deployment within organizations (e.g., Lee, 2018; Peng et al., 2023). We sought German-speaking experts from different industries and companies to provide a broad variety of practical insights. We acquired participants via LinkedIn based on their profiles. In total, we contacted 73 potential participants, of whom 33 were interested and who we provided with initial information via email. We ended up conducting 27 interviews, averaging 45 minutes in length, between March 2023 and July 2023 (mainly via Zoom or MS Teams). Table 1 is an overview of our two expert groups.

Interview design. We followed a semi-structured interview guide with three sections to ensure a standardized procedure (Myers & Newman, 2007). The semi-structured design allowed us to adjust the focus of each interview dynamically, depending on interviewees' areas of expertise and specific AI application use cases. In each interview, one researcher asked the questions and another took notes. All interviews were recorded with the consent of the interviewees and subsequently transcribed for the coding process.

The interviews were structured as follows: The first section comprised questions to elicit a description of interviewees' job activities, work culture, and the nature of their business, followed by a section about AI deployment scenarios and potential implications. The third section asked about how the experts respond to potential decision dilemmas, followed by a section of discussing ethical consideration that are required to deal with potential ethical challenges of managing human-AI collaboration in organizations, based on the work of Heyder et al. (2023). In the analysis, we explicitly distinguishing between decision dilemmas that the experts had directly encountered and those they could envision as plausible. We conducted four pilot interviews to gauge the understandability of our questions and the interview timeline. Appendix A provides the semi-structured interview guide.

Group	I D	Position	Organization	Size	Use Cases	Strategy
Leaders who decide on AI deployment (n=14)	L1	Chief Operative Officer (COO)	AI software start-up	>50 employees	Algorithmic-based predictions in marketing; AI tools for picture generation and Photoshop; financing and trading decision support agent	Exploitation (aiming to digitally transform toward e-commerce) Exploration (exploring new product offerings to drive innovation)
	L2	Chief Process and Innovation Officer (CPIO)	Automotive supplier	>30,000 employees	AI recruiting agent; autonomous manufacturing	Exploitation (core business processes); Exploration (non-core business processes)
	L3	Chief Executive Officer (CEO)	Hiring agency	<50 employees	AI recruiting agent	Exploitation (optimize HR workflows)
	L4	CEO	Ad & media agency	<50 employees	Content creation with ChatGPT and Dall-E	Exploitation, driven through external forces, AI hype (punctuated equilibrium)
	L5	Chief Human Resource Officer (CHRO)	Insurance company	>10.000 employees	Damage assessment with AI; AI recruiting agent	Exploitation (focus on improvement of core- and non-core business processes for cost-saving)
	L6	Chief Marketing Officer (CMO)	Renewable energy provider	<500 employees	AI drone operations; content creation with ChatGPT	Exploitation (core business processes); Exploration (non-core business processes, mostly unsupervised)
	L7	CEO	Industrial automation engineering company	>100 employees	AI detection in manufacturing (predictive maintenance)	Exploitation, driven by fear of replacement from other players (punctuated equilibrium)
	L8	CEO	Pharmaceuticals (previously sports equipment manufacturer)	>1,000 employees	Data analytics for performance evaluation; digital skiing teacher	Exploration (designing innovative technologies and explore data insights)
	L9	CMO	News & media platform	>2,000 employees	Content creation in journalism, predictions in marketing	Exploitation (evaluating and testing AI algorithms to optimize their output)
	L10	Lead Developer	Electronics provider	>300,000 employees	Smart home	Exploration (product innovations)
	L11	Project Lead	Vocational training institute	<100 employees	Learning records platform that supports personalized studying	Exploration (experiment with new technologies)
	L12	Account Executive	Cloud services	>500,000 employees	Fraud detection in ticket sales	Exploitation (supports customers in optimizing their processes)
	L13	AI Sales Lead	Technology provider	>300,000 employees	Coding with GitHub Copilot; AI Application based on Open-Source Transformer-Architectures	Exploration (experiments with novel AI models as competitive edge)
	L14	AI Product Lead	Medical technology services	>300,000 employees	Smart Manufacturing & Anomaly Detection through data analytics	Exploitation (leveraging data analytics to improve operational efficiency)
Employees working with AI (n=13)	E1	Software Developer	Software company	>220,000 employees	AI-based customer service in finance	Exploitation (AI-based customer support to streamline processes)
	E2	Data Scientist	Auto manufacturer	>150,000 employees	AI simulation and image generation; ChatGPT for information search	Exploration, mainly (experimenting with novel technologies for customer attraction)

						Exploitation (improving existing processes in manufacturing)
	E3	IT Security	Manufacturing company	>80,000 employees	AI bots used for cybersecurity and threat mitigation	Exploitation, driven by external forces through hackers using AI bots (punctuated equilibrium)
	E4	Recruiter using AI for C-Level Search	Headhunting agency	<10 employees	Candidate assessment with AI	Exploration (experimental use of AI tools)
	E5	Software Developer	Tax and audit consulting	>300,000 employees	Algorithmic platform for predictions; GitHub and Microsoft Copilot for programming and coding;	Exploration (latest AI trends to maintain market competitiveness)
	E6	Recruiter using AI for Candidate Sourcing and Selection	Insurance company	>30,000 employees	Candidate assessment with AI; Employer branding content creation with generative AI;	Exploitation for core business activities (supervised AI use) Exploration in uncritical tasks (unsupervised AI use)
	E7	Recruiter using AI for Candidate Sourcing and Selection	Technology provider	>135,000 employees	Preselection of candidates with AI	Internally exploitation, externally exploration
	E8	Data Analyst for Robotics	AI lab	>10 employees	Social robots in care work; robots in manufacturing	Exploration (research in the area of social robots)
	E9	Consultant for AI Tools in Education	Business analytics institute	>50 employees	Educate students on AI and governance mechanisms	Exploration (teaching through innovative learning platforms)
	E10	Data Scientist	Software company	>100 employees	Customer chatbot; coding with GitHub Copilot	Exploration (customer services); Exploitation (coding processes)
	E11	Employee at an AI Research Chair	University	>10 employees	Anomaly detection & intelligent manufacturing	Exploration (offering a new-to-market analytical tool for predictive maintenance)
	E12	Innovation Architect	Venture capital	>10,000 employees	Information search with ChatGPT; matching tool in real estate market	Exploitation (use AI to enhance existent processes)
	E13	Employee at an AI Research Chair	University	>10 employees	Mood detection through face analysis	Exploration (familiarizing students with AI technologies)

Table 1: Overview of Experts Interviewed

Data analysis. We recorded and transcribed the interviews to conduct qualitative content analysis using MAXQDA software. We mainly followed (in steps one to three) the approach of Gioia (2013) and Gioia et al. (2021) to identify stakeholder perspectives on decision dilemmas when deploying AI, as well as potential ideas on how to respond to them. The first and second authors conducted the data analysis in four distinct steps as illustrated below, following an iterative process. After each step, both authors engaged in multiple rounds of discussion to refine and validate the assignment of the codes. Figure 1 depicts the data analysis process.

In the *first* step, the authors read the empirical data, using open coding to make sense of the different use cases about AI deployment and potential dilemmas. Our approach to developing the first-order constructs was *informant-centered*, meaning we adhered closely to the language and terminology used by the interviewees (D. Gioia, 2021). This approach was designed to minimize the influence of our own worldviews on the interpretation of findings, ensuring that the study insights were firmly “predicated on the informant’s experience” (D. Gioia, 2021, p. 23). During this phase, we actively sought similarities and differences within the data that were linked to our research questions.

In the *second* step, we developed second-order themes by organizing and interpreting the first-order constructs. This phase is more appropriately described as *theory-centered* (D. Gioia, 2021, p. 23), as we sought “explanatory concepts and themes that might provide deeper insight into the informants’ experience” (D. Gioia, 2021, p. 23). While the first step focused on identifying notable findings, the second phase involved interpreting these findings by identifying themes that led us to group the first-order constructs. In the second step, we further categorized the identified quotes into two key themes, aligned with our research questions.

In the *third* step, we categorized the decision dilemmas and responses into four overarching dimensions, derived through iterative engagement with the second-order themes. These overarching dimensions, the decision-making dilemmas, also formed the basis for structuring the topics presented in our findings.

The *fourth* step, deductive mapping, aimed to identify patterns and relationships between strategies, the dilemmas they present, and their responses to address these challenges. Thus, we conducted deductive mapping

by labeling the decision dilemma identified according to the ambidexterity strategies (exploration and exploitation strategies) or punctuated equilibrium, as defined in the theoretical background.

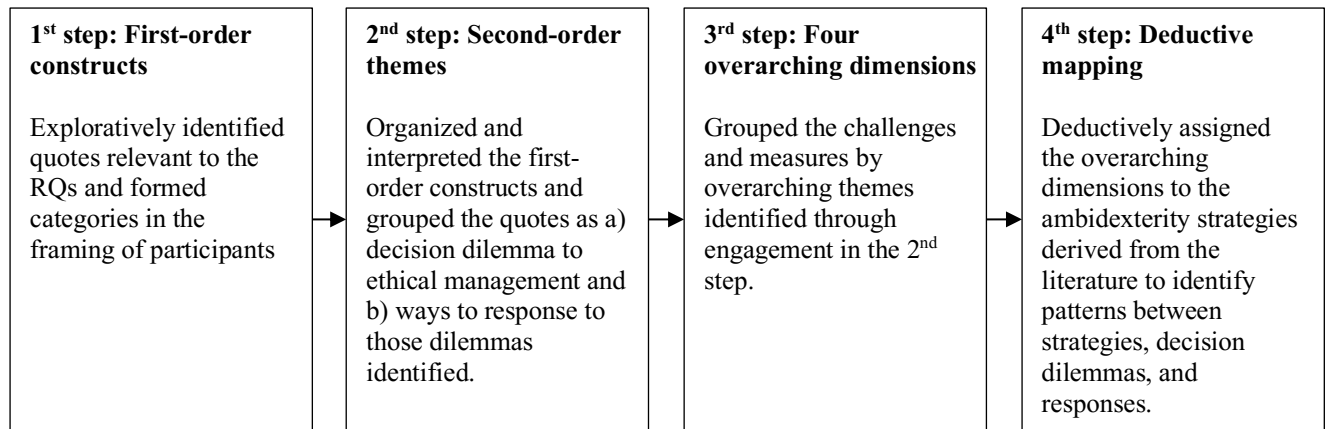


Figure 1. Data Analysis Process (Adapted From D. Gioia, 2021)

4. Results

In this section, we outline the results of our qualitative analysis. We answer our research questions regarding the decision dilemmas organizations face within AI deployment strategies (RQ1), and how they respond to potential dilemmas that may arise (RQ2). In the following, we outline four main decision dilemmas – namely:

1. Balancing a culture of shared responsibility with clear allocation of individual responsibility
(mainly within *exploration* strategies)
2. Balancing the effective use of AI's capabilities with the risk of ethical violations
(mainly within *exploitation* strategies, strengthened by *punctuated equilibrium*)
3. Balancing standardized processes with the potential for personalized approaches
(mainly within *exploitation* strategies, strengthened by *punctuated equilibrium*)
4. Balancing trust in AI with the risk of blind overtrust
(mainly within *exploration* strategies)

and illustrate how our interviewees' organizations respond to them. See Appendix B for a summary of the four decision dilemmas dependent on the different types of tasks of the AI deployment context.

#1 Balancing a culture of shared responsibility with the clear allocation of individual responsibility

A dilemma particularly pronounced in *exploration* strategies, where organizations are testing novel and opaque AI technologies such as predictive simulation or AI-assisted healthcare robots (e.g., E2, E8), concerns the need to weigh a clear allocation of responsibilities against a corporate culture that is oriented to shared responsibilities. While clear responsibility allocation would help clarify accountability for (potentially negative) outcomes of human-AI interaction, a culture of shared responsibilities fosters a willingness among all stakeholders involved in human-AI interaction to take responsibility. The contexts reported in which this dilemma occurs often lack transparency, making it challenging for several stakeholders (e.g., designers, users, or leaders) to fully understand and take responsibility for the outcomes of AI interactions. Moreover, exploration's iterative and collaborative nature was said to complicate the establishment of clear accountability frameworks, as exploration relies on flexibility and shared learning among diverse stakeholders. Our exploratory analysis identified various reasons that underlie this dilemma:

First, this decision dilemma arises when AI tools absolve users of the responsibility for decisions because those users lack a sophisticated understanding of AI's capabilities (e.g., E7, L8, L14). For instance, one leader argued that customers “*rarely question how [AI recommendations] are generated. The art lies in creating solutions so user-friendly that technical knowledge becomes unnecessary*” (L14, p.5). One of our recruiters illustrated a potential consequence of AI's non-transparency, stating that recruiters might “*hide behind AI's decisions – especially when they do not have an idea of how the AI operates. This results in justifying the AI's rejection of applications by stating ‘we can't do anything; the system said so’*” (E7, p. 10). A leader who described how her company's culture changed through AI integration provided another example, explaining that prior to AI implementation people were held personally responsible, but with AI it is difficult to allocate responsibility because there are too many stakeholders involved before (e.g., data scientists providing data input) both during (e.g., developers designing AI's parameters) and after (e.g., managers working with AI suggestions) the interaction with AI (L8). Thus, the dilemma of prioritizing a culture of shared responsibility over clear responsibility allocation may stem from AI's inability to fully comprehend contextual nuances, which shifts responsibility away from individual actors to a diffuse network of stakeholders.

Second, this dilemma arises from a natural human tendency to seek recognition for successes while deflecting blame or finding excuses for failures – a behavior that can be extrapolated to human-AI collaboration. For instance, L10 observed that employees often prefer to blame colleagues rather than AI systems, while L4 suggested that it is easier to attribute fault to AI technologies because they cannot defend themselves against accusations, making them convenient scapegoats. A C-level manager (L6) added that he is indifferent to whether results are achieved with or without the assistance of AI tools if the outcomes are accurate. He posited that, similar to journalists being held responsible for their written content rather than the opinions of their audience, individuals remain responsible for the final output regardless of AI involvement. However, the context of healthcare applications, where robots administer medication (L11), responsibility was attributed to the organization deploying the AI as with increasing level of automation, responsibility shifts away from single users. This discussion highlights that the dilemma is influenced by societal concerns about reputation loss and the degree of human intervention allowed by the AI, as its level of autonomy directly influences the allocation of responsibility.

To respond to the decision dilemma inherent in exploration strategies—namely, how to balance shared responsibilities with clearly defined individual accountability—our qualitative analysis suggests that traditional concepts of *structural ambidexterity* may require extension. Specifically, rather than viewing ambidexterity as a process confined to the alignment of exploration and exploitation across internal organizational units, our findings point toward the need for vertical integration of diverse external stakeholders—including industry experts, regulatory bodies, and public institutions—into strategic AI deployment processes. For example, policymakers often focus on exploitation by refining and enforcing existing regulatory frameworks for AI (e.g., L2, L13, E9). In contrast, industry actors tend to adopt a dual strategy: on the one hand, developing and enhancing practical guidelines for the responsible use of AI; on the other, pursuing innovation to address novel challenges and opportunities in unexplored domains (e.g., E1, L5, L10, L13). As one interviewee emphasized, the establishment of an industrial ethics committee could help define minimum ethical standards for AI applications in the private sector (L11). Similarly, another participant proposed the development of a policy reference framework to guide ethical AI deployment in practice (L6). Organizations engaged in exploratory strategies frequently reported the need to mitigate the inherent uncertainty of innovation by involving a broad

range of stakeholders early in the development process. For such collaborative innovation to be effective, individuals across the ecosystem must clearly understand their roles and accompanying responsibilities (e.g., L6, L9). As one leader put it: *“I advocate for personal responsibility, both for myself and others. When it comes to AI, I see it as a tool. Whether I succeed or fail, I don’t blame the tool; I recognize that ultimately, it’s the person using it who bears responsibility for the outcome”* (L6, p. 7). This perspective illustrates how leadership behavior – particularly when leaders act as ethical role models – can foster a culture of shared responsibility. Such a culture, in turn, supports ethical decision-making across various stakeholder groups involved in human–AI collaboration, including developers, data scientists, managers, and end users (e.g., L12, L14).

These insights suggest that resolving the dilemma of shared versus individual responsibility in AI-driven exploration may require a collaborative approach to structural ambidexterity that goes beyond intra-organizational coordination. Rather than being limited to the sequencing of exploration and exploitation within a single organization—as traditionally conceptualized—this expanded view of structural ambidexterity involves a coordinated, multi-level strategy that includes external stakeholders.

#2 Balancing the effective use of AI’s capabilities with the risk of ethical violations

In the context of *exploitation* strategies, our interviews revealed that organizations sometimes face decision dilemmas that concern the potential for abuse when deploying AI despite AI’s potential for increasing the efficiency of processes. This dilemma emphasizes the dual nature of AI technologies, which can be wielded for both beneficial and harmful purposes. At times, it is externally driven by disruptive forces (punctuated equilibrium), such as cyberattacks employing AI bots, which compel organizations to deploy AI defensively in response. It is important to note that AI-providing organizations do not choose between these outcomes as part of their strategy but instead recognize that both appropriate use and misuse can arise among users trying to leverage AI’s capabilities. The dilemma may be present in different scenarios, driven by various underlying reasons:

First, the design of AI technologies often enables unintended outcomes, as AI technologies cannot inherently discern the intent behind a user’s prompt or action. For example, one interviewee highlighted how a generative AI tool, such as ChatGPT, could respond to a prompt about breaking into a car (L11). He argued that

the technology cannot differentiate between malicious intent, as in stealing the car, and altruistic intent, such as saving a child locked inside the car on a hot summer day (L11). Similarly, generative AI tools may be employed to enhance social media content or propagate misinformation (e.g., L1, E13). Another interviewee offered the example of monitoring tools, which can be used to reward employee performance but may also violate worker privacy by tracking behavior too intrusively (e.g., L12), such as by monitoring mouse movements or automatically logging workers out of their accounts for inactivity (E6). Thus, similar to previous technologies or all other forms of material entities, AI technologies have a dual nature and can be misused. One interviewee framed it: “[with] a chainsaw you can cut down trees, which would be the actual purpose, but you can also do a lot of damage with it – similarly with AI” (L6). An unethical use can even happen unconsciously. For instance, one leader noted that sometimes their employees upload sensitive information to ChatGPT without being aware that the generative AI will use that information to train their models (L4).

A *second* reason for this dilemma lies in AI’s potential to extend both altruistic applications and personal gains. For instance, one recruiter reported that they might contrive AI to maximize outcomes for others (such as family and friends; E7, E6). Conversely, job candidates may abuse AI for their own benefit (such as seeking to gain an unfair advantage by tailoring Curriculum Vita’s (CVs) to align with what the AI prioritizes; E4). One recruiter argued that when testing an AI agent for job candidate search and selection, the issue is that “*candidates sometimes list skills in their application that they do not actually possess, and the AI cannot detect this. If I want to determine whether someone truly has certain professional experience, I need to address this during an in-person interview ... Of course, you can’t rely on the AI’s suggestion; you can’t catch fake profiles unless you talk to the candidates*” (E4, p. 8). A leader stated that “*to be honest, I sometimes choose not to disclose AI’s involvement when generating certain content, to avoid being criticized for using support tools. ... I think that’s fine as long as you can stand behind the work and take responsibility for it*” (L11, p. 10). In contrast, one interviewee outlined that AI could help in detecting plagiarism more effectively, thereby mitigating abuse in academic or professional contexts (L10). This highlights how AI’s capabilities can create opportunities for both ethical advancement and misuse. Particularly, loopholes in AI systems (e.g., gaps or ambiguities in the design or implementation that can be exploited in unintended ways) often create conflicts of interest that are difficult to resolve and thus can exacerbate this dilemma. For instance, one interviewee argued “*exploiting pricing errors*

in an online forum to take advantage of an advertising error for a €1 notebook instead of its correct price of €1,900. I would not post it in a forum; I would use a feedback formula to report the mistake to the company, because there is a person who made the mistake and who might have to take responsibility for it, and I don't want to make things worse by spreading the mistake" (L3, p. 18). She argued that failing to report such errors and instead profiting from them is ethically questionable, as someone ultimately bears the cost of the mistake (L3).

To respond to this dilemma ethically, our qualitative analysis outlines that traditional concepts of *sequential ambidexterity* may require extension by demonstrating how organizations can embed short cycles of ethical exploration within their longer periods of exploitation strategy to use AI effectively. To avoid ethical pitfalls when leveraging AI's capabilities, organizations may need to establish a threshold for acceptable ethical behavior, grounded in a shared understanding of what constitutes morally permissible actions, defined by a dynamic process of ethical dialogue among diverse internal and external stakeholders—including developers, managers, end-users, regulators, and affected communities (e.g., E2, E10, E11, E12). These stakeholders each engage with AI technologies through their own strategic orientations, whether exploration- or exploitation-driven, and contribute varying ethical expectations. Thus, ethical considerations are integrated not by shifting the organization's exploitation strategy, but by creating temporary and sequential integration of ethical exploration in exploitation—triggered by internal or external ethical concerns—and allowing for reflection, dialogue, and adjustment before resuming operational focus. For instance, legal and compliance stakeholders (e.g., L2) may engage early in development (short exploration input), while end-users (e.g., L12) may surface ethical issues only during experimenting with the technology (exploration input). One interviewee noted that organizations “*may opt to incentivize the identification and reporting of loopholes or misuse by offering rewards to employees or users who do so openly*” (E7, p. 13), as opposed to “*implementing overly restrictive control mechanisms that fail to address underlying issues*” (E8, p. 12). Hence, making the exploration of loopholes to an attractive goal for employees and end-users in the development phase of an AI lifecycle can be a short period of ethical exploration providing relevant ethical considerations for the organization's strategy. Recognizing and sequencing these ethical considerations enhances the organization's capacity to sustain exploitation strategies without overlooking evolving ethical pitfalls.

These insights suggest that responding to the dilemma of leveraging AI's capabilities for effective use versus avoiding ethical pitfalls may require both externally and internally-sourced exploratory periods to introduce ethical considerations into an ongoing exploitation strategy itself. Unlike traditional models of sequential ambidexterity, which emphasize temporal shifts between internally driven exploration and exploitation, this expanded view highlights a temporal stakeholder-driven integration of ethical exploration.

#3 Balancing standardized processes with the potential for personalized approaches

Our study further revealed a decision dilemma that appears particularly within strategies of *exploitation*, where the focus is on optimizing established processes and enhancing efficiency – which often requires a certain degree of standardization. Thereby, standardization as exploitation strategy is increasingly shaped by external disruptions (punctuated equilibrium), including media hype around generative AI and competitive pressures driving AI-based automation for efficiency gains. These forces push organizations to adopt standardized approaches in areas like content creation, programming, and process automation in manufacturing. The resulting dilemma lies in balancing the implementation of AI-driven standardized processes—rooted in the logic and limitations of the technology—against the need to preserve personalized, context-sensitive practices that reflect human uniqueness. Our analysis suggests that this decision dilemma may arise in different settings and due to different reasons:

First, this dilemma may arise from uncertainties about individual competencies when human decision-making comes into direct competition with AI-generated suggestions. Particularly, AI augmentation of tasks may increase performance but also lead to a role shift and reskilling of employees, potentially provoking “*identity threats among affected individuals due to organizational restructuring or reskilling initiatives ... resulting in significant uncertainties for employees whose job profiles are affected*” (E13, p. 14). For example, in damage claim assessment, AI tools equipped with predictive capabilities can accurately estimate repair costs, streamlining the work process (L5). But while this boosts efficiency, employees in that role often feel their judgment and expertise are undervalued, as key aspects of their decision-making are handled by AI (L5). Similarly, in solar park maintenance, AI drones have changed manual inspections to detect faulty modules (L6). Workers who previously drove through the fields daily to identify issues now have to fly drones, and they

intervene only when AI signals a problem; this reactive role has left some workers questioning their purpose and professional significance (L6). Identity threats may arise as one employee stated: “[when] you are used to conducting negotiations and suddenly are no longer required, you lose some of your own identity” (E7, p. 9). This shift can lead to stress when subjective perceptions of expertise conflict with AI recommendations (L9, L12). However, human decision may be emotionally biased (e.g., L7, L13). As one recruiter argued, “Through AI, the evaluation of applications would be more objective. You don’t look at the picture, you don’t look at the last name, but you go only by qualifications, for example. And I think that in this respect, more objectivity might be a good thing.” (E6, p. 10). Another recruiter argued that whether AI’s neutral decision is required compared to people’s empathic-based selection may depend on the context: “For certain jobs, such as kindergarten teachers, you need the personal evaluation, as the AI cannot pick up empathy and social skills; however, an AI tool can make the assessment when you need certain mathematical skills” (E6, p. 11). These examples show standardized approaches may increase objective assessments but to not account for human individuality.

Second, this dilemma appears when comparing unique and creative outputs people generate with homogenous outputs produced by AI. As one data scientist explained, “If every company creates its newsletter with ChatGPT, they will be very similar, generic, and the variety will be reduced” (D1, p. 3). This homogenization has raised concerns among media companies, particularly regarding potential negative impacts on their Google rankings when algorithms detect generative AI content (e.g., L9). Such concerns are especially pronounced in areas such as employer branding and marketing, where content must stand out, create certain feelings, and demonstrate originality (L4, L9). However, as a marketing agency CEO shared: advances in generative AI tools have led to the “dismissal or retraining of existing copywriters” (L4, p. 3) given that customer relucted to pay for full-time copywriting services in light of freely available AI tools. Hence, this dilemma is particularly significant for businesses in which creativity and innovation form the core of their business models, such as media and advertising agencies. Likewise, in recruitment processes, the use of AI for candidate selection can expedite standardized procedures but may inadvertently convey a sense of devaluation among individual candidates: “If there are artificial interviewers that take only ten minutes, candidates quit [the application process] because they no longer feel valued” (E4, p. 3). However, one manufacturing CEO argued that their growing dependency on market standard software that is increasingly based on AI also forces them to

establish such standards and reduce customization (L7). Similarly, in the context of data management, the service ecosystems of global providers do not allow for individual software packages, so companies have to implement a standard portfolio of external providers, creating certain dependencies (E11). One leader highlighted that unforeseen circumstances might necessitate human intervention to address unique challenges, particularly in exceptional or accidental scenarios where rigid AI-generated responses might fail to capture the nuances required for appropriate action (L6). Thus, while AI may excel in making rational and standardized decisions, it may not always be suitable for dynamic and complex scenarios requiring adaptability and creative problem-solving (L10, L11).

To respond to the dilemma of weighing decisions that prioritize standardized processes against personalized approaches—predominantly within exploitation strategies—our findings suggest the need to extend the traditional concept of *structural ambidexterity*. Typically, structural ambidexterity assumes that exploration and exploitation are best managed through the separation of units—one focusing on innovation and experimentation, the other on efficiency and refinement. However, our qualitative analysis reveals a more nuanced reality: tensions between standardization (exploitation) and individualization (exploration) often emerge within the same role, team, or task—rather than across clearly divided units. For example, an HR professional may rely on AI-based recruitment tools to ensure fairness and objectivity in candidate assessments (a standardized, exploitative approach), while simultaneously crafting individualized job advertisements that aim to emotionally engage and resonate with specific audiences (an exploratory, creative task) (L3, L9). This demonstrates that the same individual or unit may need to balance between exploration and exploitation, depending on the context and ethical implications of a given decision. Such intra-role ambidexterity challenges the binary logic of structural separation and instead calls for new organizational mechanisms that support ethical reflection and decision-making across functional boundaries. Furthermore, in multinational organizations, this ethical complexity is amplified by cultural and moral diversity across locations – as interviewees noted significant differences in moral standards between Western countries and regions like Russia or China (e.g., E9, L5). As one leader put it, “*there is no objective fairness for everyone*” (L5, p. 15), highlighting the challenge of defining ethically acceptable standardization for exploitation strategies across diverse settings.

These insights suggest that resolving the dilemma of standardization versus personalized approaches within exploitation strategies may require an extension of structural ambidexterity beyond the separation of exploration and exploitation units. Rather than relying solely on structural separation, organizations must enable alignment across diverse departments and cultural contexts—embedding mechanisms for inter-unit dialogue and shared ethical reflection to ensure that operational efficiency through standardization does not come at the expense of contextual sensitivity and human-centered values.

#4 Balancing trust in AI with the risk of blind overtrust

Concerning *exploration* strategies involving the use of AI to identify new markets and develop innovative business ideas, our interviewees reported encountering decision dilemmas linked to trust. These dilemmas arise from the need to balance fostering employee trust in AI technologies with preventing overtrust. Maintaining a healthy level of skepticism and critically evaluating AI outputs is essential for ensuring ethical interaction with these systems. According to our interview findings, this dilemma becomes particularly relevant in situations where individuals either lack domain knowledge or anthropomorphize AI – meaning a human-looking design of the AI technology and its ability to imitate human behavior. However, the potential explanations for what triggers the dilemma varied across cases and AI technologies:

First, AI's increased anthropomorphic AI design is feared to aggravate this dilemma (e.g., L8, E8). Thus, exploration strategies concerning the implementation of robots in social contexts revolve around weighing decisions that should cultivate individual trust in AI to foster adoption through an anthropomorphizing design of AI technologies against decisions that mitigate the risk of anthropomorphism to avoid disrupting social dynamics among workers or potential clients. Our qualitative analysis revealed that this decision dilemma often arises in scenarios where AI replaces human roles and mimics human behavior. One leader shared the development of a digital ski instructor that uses sensors attached to ski boots and an AI-powered smartphone app for training purposes (L8). Based on data from professional skiers, the app provides tailored advice designed to build user trust and drive purchases – “*advice [that] should be deemed very appropriate,*” (L8, p. 5) and to which “*users might then attribute some human intelligence*” (L8, p. 5). However, during its introduction, the project team expressed significant concerns that customers might blindly follow the AI's instructions, which

could result in accidents given that the digital ski instructor cannot account for factors such as whether students are fatigued and need rest (L8). Another example included insights from the use of AI robots in healthcare outlined that AI robots emulating human social behavior may lead patients to perceive them as genuinely empathetic since “*robots respond to emotions, some patients might think they really are empathic. So, what happens if patients start to rely too heavily on the robot and perhaps lose social contacts?*” (E8, p. 6). While anthropomorphism may help patients trust AI robots and convince them to follow their instructions to take their pills, it may also lead to patients overly relying on them, potentially leading to social isolation (e.g., L8, L11). Thus, trust and overtrust may be aggravated by anthropomorphic design and the context of AI deployment.

Second, AI tools’ increasing usability was reported to offer a broad mass of people access to areas where they previously had no domain expertise (e.g., D4, D5, L4, E5, E6). While this represents a significant advantage of AI tools, gaining access to areas without prior domain knowledge can, in the worst case, lead to dependency effects and overreliance. For example, as one expert explained, code becomes more creative and precise when employees use CoPilot, and overall efficiency improves, particularly for individuals who are new to programming (E6). However, “*the real danger is that someone with no knowledge can use AI to do certain things without having the necessary training*” (L10, p. 3). As a consequence, people sometimes lack the ability to question outcomes, which may result in a forced overreliance on AI (e.g., L8); as one interviewee put it, “*understanding AI algorithms isn’t exactly straightforward. You need to be familiar with statistical methods*” (L8, p. 5). Another leader highlighted the unsupervised use of ChatGPT for research by the workforce within the organization (L4). While the practice was actively encouraged and supported, the leader recognized the risk that AI-generated outcomes may not be critically evaluated or scrutinized (L4). Thus, trust and overtrust may be aggravated by very diverse understandings of what AI is and what it can do (e.g., E13, L12, E1).

To respond to the dilemma of balancing trust in AI with the risk of overtrust—primarily emerging within exploration strategies aimed at leveraging AI for innovation and new opportunities—our findings suggest the need to extend the concept of *sequential ambidexterity*. Traditionally, sequential ambidexterity implies alternating between phases of exploration and exploitation across time. However, our qualitative data reveals that within extended periods of exploration, temporary top-down exploitative interventions may be necessary to mitigate ethical risks such as overreliance on AI or misuse due to limited domain knowledge. For instance,

while organizations encourage exploratory use of generative AI tools to stimulate creativity or enhance productivity (e.g., L4), many also recognize the need for structured interventions—such as critical reflection, standard-setting, and AI literacy training—to help employees differentiate between mimicry and genuine intelligence (e.g., E5, E9). These exploitative interventions serve to stabilize exploration by fostering healthy level of critical awareness and ethical standards in working procedures. As one interviewee put it, “*It takes more training and planned personal conversations among workers to examine AI’s limits*” (E7, p. 6), while another noted that transparency around AI’s decisions and outputs can help prevent blind trust (E2, L2).

These insights suggest that resolving the dilemma of trust versus overtrust within exploration strategies may require an extension of sequential ambidexterity – one that integrates short, top-down exploitative interventions into ongoing exploratory efforts, ensuring that trust in AI remains ethically grounded by optimizing existing internal resources such as employees’ AI literacy and ethical standards embedded in working procedures.

5. Discussion

In this section, we discuss the theoretical and practical implications of our findings. We also outline its limitations and offer recommendations for future work.

5.1 Theoretical Implications

Building on a growing body of research that emphasizes the importance of ethical reflection in organizational decision-making around AI deployment (e.g., Berente et al., 2019; Ling et al., 2022; Mirbabaie, Brendel, et al., 2022; Shahid et al., 2023), our study demonstrates that ethical considerations are not merely a regulatory concern but are central to achieving ambidexterity in practice. Thereby, our theoretical research contributions are twofold:

First, we identify four main decision dilemmas that arise when AI is deployed in organizational contexts and contribute to the ethical discourse by proposing a perspective that extends beyond setting up ethical principles and compliance-based norms (closely linked to deontology). Specifically, we advocate for a more holistic ethical approach by examining the values and behaviors of diverse human stakeholders within organizations, thereby, incorporating a virtue ethical perspective on decision dilemmas of AI deployment. This

perspective is crucial for addressing and managing human–AI collaboration ethically (Heyder, 2023), and directly responds to recent calls in the IS literature to advance AI ethics discourse (Hafner et al., 2024; Hammerschmidt et al., 2025; Mirbabaie et al., 2022).

Second, by examining how organizations navigate the dual demands of exploration and exploitation in real-world AI deployment scenarios, we provide ideas on how to extend current concepts of ambidexterity to integrate ethical considerations. In doing so, we address recent calls in the IS literature to more explicitly integrate ethics in ambidexterity research (e.g., Ling et al., 2022; Shahid et al. 2023; and Yang et al. (2024)).

Expanding AI ethics discourse with ethical decision dilemmas of AI deployment

In the context of managing AI deployment, the prevailing ethics discourse remains largely focused on establishing policies and regulations (Hafner et al., 2024, Hammerschmidt et al., 2024). This emphasis reflects a primarily deontological perspective – centered on norms and duties – which does not fully capture the breadth of ethical considerations necessary for a holistic understanding of ethical AI deployment in organizations. In line with prior research (e.g., Berente et al., 2021; Hafner et al., 2024; Mirbabaie et al., 2022), we argue that more empirical insights are needed into how the inherent moral values, ethical perceptions, and behaviors of diverse organizational stakeholders, closely aligned with a virtue ethics perspective that influence the ethical outcomes of human–AI collaboration (Heyder et al., 2023). Our study contributes to this discussion by drawing on practical insights to inform theory, identifying four ethical decision dilemmas that arise in AI deployment through the lens of participants’ virtue ethics, and thereby advancing the discourse on AI ethics in IS research.

The first decision dilemma concerns the tension organizations face the tension between shared responsibility and clear allocation of individual responsibility. This tension especially appears with emerging AI applications that are often adopted in a bottom-up manner by employees during experimental phases of exploration. The absence of formal regulations (reflecting a deontological perspective) increases the risk of unclear ownership over negative or unintended outcomes. Consequently, fostering a culture in which individual organizational stakeholders are willing to take responsibilities (a virtue ethics approach) is critical. Our data suggests that both approaches are important: clear governance structures (e.g., E13) and the cultivation of moral compasses shaped

by organizational stakeholders' cultural and social backgrounds (e.g., E9, L15), thereby, reflecting their virtue ethics.

The second decision dilemma concerns the trade-off between leveraging AI's capabilities for effective use and avoiding ethical violations. While AI holds substantial potential in sensitive areas such as cybersecurity, these contexts also entail increased risks of misuse and harm. Misuse may be unintentional, but there is also the risk of intentional abuse by organizational stakeholders who lack a virtue ethics orientation—for example, by exploiting loopholes (e.g., E2, E7) or engaging in harmful hacking activities (e.g., E3). Even when AI is deployed by organizations to optimize existing processes (i.e., exploitation) and supported by formal guidelines and instructions (reflecting a deontological approach), organizations have limited control over the ethical disposition of individual stakeholders. This limitation was also evident in our empirical findings, where many leaders emphasized that ethical outcomes increasingly depend on individual personality and moral values (virtue ethics), particularly in hiring decisions (e.g., L2, L3, L6).

The third decision dilemma highlights the tension organizations face between standardization and personalized approaches during exploitation-oriented strategies. For instance, in AI-supported content generation, a focus on efficiency may conflict with the need for creativity, uniqueness, and contextual sensitivity (e.g., D1, L4, L9). From an ethical perspective, prioritizing standardization through usage guidelines or predefined policies (aligned with deontological ethics) can risk marginalizing individual needs, limiting diverse human perspectives, and reducing inclusivity (e.g., E6, L7), thus, conflicting with a virtue ethics approach that would emphasize organizational stakeholders' moral character and context-sensitive judgment.

The fourth decision dilemma addresses the challenge of fostering trust in AI technologies without encouraging blind overtrust. While building confidence in novel AI systems is essential for their adoption, excessive reliance may undermine critical human oversight. This dilemma is particularly evident in exploration strategies involving anthropomorphic AI technologies or AI robots performing physical tasks (e.g., E8, L8). It requires striking a balance between promoting user trust and implementing safeguards that preserve human judgment. From an ethical standpoint, this necessitates a combination of deontological principles—to ensure oversight mechanisms are in place—and virtue ethics, which guide moral reasoning about what is considered

acceptable. Our data suggest that the design of such technologies should incorporate diverse virtue ethics perspectives to assess how AI agency is framed and to anticipate potential ethical concerns that may conflict with individual moral values (e.g., E2, E9). Ethical committees with clear guidelines or advisory structures (e.g., L11, L13) were mentioned as potential mechanisms to support this process.

Together, these decision dilemmas demonstrate that balancing exploitation and exploration in AI deployment is not only a strategic challenge but also a fundamentally ethical one. Consistent with prior research (e.g., Bonnefon et al., 2024; Kegel & Ghanem, 2024), these dilemmas often involve situations where any chosen course of action may lead to adverse consequences for certain organizational stakeholders, thereby increasing the risk of unintended negative outcomes. As discussed in the theoretical background (Section 2.1), decision dilemmas in ambidextrous strategies are inherently ethical, as they require aligning strategic objectives with ethical considerations. This underscores the need for a multifaceted ethical approach, rather than a one-sided, compliance-oriented perspective, when addressing ethical concerns in AI deployment strategies. Drawing on qualitative insights, our study enriching the AI ethics discourse with a virtue ethics perspective and highlighting that organizational ambidexterity in the AI era must integrate ethical reasoning rooted in virtue ethics, beyond the traditional trade-off between efficiency (linked to exploitation) and innovation (linked to exploration).

Extending ambidexterity concepts by considering ethical tensions

In line with recent research on technology-driven exploration and exploitation strategies (Wang et al., 2024), we argue that effectively addressing the complexities of ambidexterity in AI contexts requires a theoretical extension of existing ambidexterity frameworks. Our empirical insights suggest that traditional concepts—which emphasize structural or sequential ambidexterity—must be extended to account for the ethical considerations to respond to decision dilemmas that arise within AI deployment.

Ideas on how to extend structural ambidexterity concepts to address ethical tensions in AI deployment

Structural ambidexterity traditionally assumes the separation of organizational units or teams that pursue either exploration or exploitation, allowing each to focus on divergent goals while maintaining overall strategic alignment (Cao et al., 2024; De Visser et al., 2010). Typically, innovation and experimentation are delegated to

dedicated R&D or innovation teams, while core business units focus on efficiency, refinement, and incremental improvements (Cao et al., 2024). Coordination between these units is maintained through top-level integration, allowing the organization to pursue both strategic logics without internal conflict (e.g., De Visser et al., 2010). However, our findings—particularly those related to the first and third decision dilemmas—reveal that such separation is insufficient for managing the ethical complexities of AI deployment. Based on our data, we propose two complementary extensions of structural ambidexterity that better account for these challenges:

First, a vertical extension with a multi-level, stakeholder-oriented approach to structural ambidexterity. While traditional structural ambidexterity focuses on intra-organizational coordination, our findings highlight the need for a vertical extension that consider external stakeholders into ambidextrous decision-making processes. Our data illustrate that in the context of AI, stakeholders such as regulators, civil society actors, industry associations, and even end users play a critical role in shaping both opportunities and risks of AI (e.g., E9, L2, L13). For example, ethical concerns about algorithmic bias may not originate internally but are raised by external actors, requiring organizations to coordinate exploration activities (e.g., assessing long-term ethical risks) and exploitation strategies (e.g., operationalizing fairness standards) beyond their organizational boundaries (e.g., E7). This broader, multi-level and stakeholder-oriented approach to structural ambidexterity emphasizes ethical co-design and co-regulation as core mechanisms to respond to decision dilemmas of AI deployment ethically.

Second, an intra-role extension of structural ambidexterity. Traditional concepts assume that exploration and exploitation are managed through the functional separation of units. Yet, our data suggest that decision dilemmas emerge within the same role, team, or task, and so do and must exploration and exploitation strategies. For instance, employees in HR or marketing may operate under primarily exploitative logics—such as standardizing recruitment or automating content generation—while simultaneously performing exploratory tasks that require human sensitivity, creativity, and contextual judgment (e.g., E4, L4). A recruiter may rely on standardized, AI-driven assessments to promote fairness, yet also customize job postings to attract diverse applicants (e.g., E6, L5). These intra-role tensions are intensified in multinational organizations, where cultural and moral values differ across locations (e.g., E9, L5), making centralized structural separation ineffective. To

manage these tensions ethically, organizations must support cross-functional dialogue, embed reflective practices, and create spaces where individuals can reconcile competing demands within their daily work.

Ideas on how to extend sequential ambidexterity concepts to address ethical tensions in AI deployment

Sequential ambidexterity traditionally refers to organizations alternating between distinct phases of exploration and exploitation over time, typically managed by a single organizational unit or entity (e.g., Goossen et al., 2012). This temporal separation allows organizations to focus on innovation during exploratory phases and efficiency during exploitative phases, maintaining overall strategic balance. However, our empirical findings—particularly relating to the second and fourth decision dilemmas—suggest responding to potential decision dilemmas requires extensions to adequately address the ethical complexities inherent in AI deployment:

First, a top-down exploitative intervention within exploration phases is proposed as one possible extension of sequential ambidexterity. Our qualitative data reveal that extended exploratory periods—such as experimenting with novel AI applications—cannot rely solely on open-ended innovation. Instead, temporary, targeted exploitative interventions initiated from top management are often necessary to mitigate ethical risks, including overreliance on AI or misuse stemming from limited domain knowledge (e.g., E7, L10). For example, organizations implement short-term measures such as AI literacy training, ethical standards in workflows, or critical reflection sessions during ongoing experimentation with generative AI tools (e.g., L8, L10). This integration of brief, top-down exploitation ensures that trust in AI remains well-calibrated and ethically grounded throughout exploration. Thus, we propose an extension of sequential ambidexterity that embeds exploitative ethical considerations and balances within longer exploratory cycles—optimizing internal resources to ensure an ethical AI deployment. This is particularly relevant when external forces trigger rapid shifts in exploration – an idea consistent with punctuated equilibrium (Gupta et al., 2006, O'Reilly & Tushman, 2004). For instance, when one firm launches innovative AI solutions, competitors may be pushed into reactive exploitation strategies to maintain their market position, such as automating processes to cut costs.

Second, a bottom-up intervention with cycles of ethical exploration within periods of exploitation. Conversely, traditional models frame sequential ambidexterity as a clear temporal shift from exploration to

exploitation. Our findings add another value to this view by showing that during extended exploitation phases, characterized by scaling and refining AI-driven processes, organizations must actively embed short, periodic cycles of ethical exploration that should be driven from the bottom-up. These cycles serve to re-examine ethical implications and integrate stakeholders' ethical considerations to ensure ongoing ethical AI deployment. This extension becomes particularly important in situations marked by externally triggered disruptions that appear bottom-up, such as rapid shifts in public expectations (e.g., L4)—a phenomena often described by the punctuated equilibrium forcing organizational change (Gupta et al., 2006, O'Reilly & Tushman, 2004). By embedding stakeholder-driven cycles of ethical exploration into periods of exploitation, organizations can adapt more responsively to ethical risks without abandoning efficiency goals.

In conclusion, our findings expand the ambidexterity discourse with some ideas on possible extensions to applying ethical considerations when facing decision dilemmas within AI deployment strategies.

5.2 Practical Implications

Our study addresses how organizations try to respond ethically to decision dilemmas within their exploration and exploitation strategies of AI deployment, and outlines how organizations may manage human-AI collaborations with respect to ethical considerations, as requested by Alsheiabni et al. (2020), Mirbabaie et al. (2022), Vial (2019) Yang et al. (2024), and Zhang et al. (2022).

Our empirical data outline that those in our leader sample reflect more about actual decision dilemmas, while employees tend to report more on measures to address those dilemmas. This difference leads to distinct strategies by internal stakeholders responding to ambidextrous decision dilemmas within organizations. Employees, as the group most directly impacted by the unintended consequences of AI deployment, often exhibit a heightened awareness of the implications of AI deployment and advocate for measures to safeguard themselves from negative consequences. In our survey, employees expressed a greater need than leaders for enhanced transparency regarding the underlying parameters of AI technologies with which they interact. This emphasis on transparency and explainability stems from employee concerns about dependencies on AI providers (e.g., E3) and preventing potential unintended misuse because of a lack of background knowledge (e.g., E11). Leaders, though, who typically operate at a more strategic level and might less engage directly with

AI technologies, tend to report a lower urgency for such transparency measures. One leader in our sample even argued for measures in the opposite direction, stating that given the fundamental lack of openness to technology in the particular company, less transparency within certain technologies would allow employees to gain initial experiences and establish trust in AI before it was made transparent (L8). Thus, this study suggests that practitioners should deliberate about how much transparency will benefit AI deployment strategies by reducing hierarchies for deliberation on equal terms.

Finally, our findings reveal that the reported measures on the organizational level are most often linked to exploration, while those on other levels (e.g., industrial code of conduct or societal-driven AI regulation) are often related to exploitation strategies. While this may reflect that our sample focuses on the internal stakeholders of organizations, there may be an actual imbalance of exploration and exploitation between the organizational, industrial, and societal levels. This suggests the need for an alignment of strategies, processes, and resources that go beyond organizational borders, supporting the arguments of Constantiou et al. (2023) and Bonnefon et al. (2024). For instance, organizations may seek opportunities for synergistic collaborations and an interdisciplinary dialogue (e.g., E12, L12); collaborative ventures with industry partners, academia, or governmental institutions can facilitate knowledge sharing, resource pooling, and the co-creation of value. These could help organizations develop ideas for how to incorporate ethical considerations into their AI strategies. In addition, organizations could proactively establish ethical principles that exceed existing regulations. This is supported by our qualitative interviews, in which leaders argued that existing regulatory frameworks – perceived as geographically overspread and too vague for concrete orientations – often fail to provide clear guidance regarding morally acceptable thresholds (e.g., L1, L4, L12), which underscores the need for industry-led initiatives to fill this gap.

5.3 Limitations and Future Work

Our work has several limitations that also point to future research directions.

First, our findings are based solely on internal organizational stakeholders, limiting insights into the role of external actors, such as policymakers, educators, customers, and business partners. These stakeholders may significantly influence AI deployment strategies and ethical considerations – as outlined in our theoretical

implications on how to extend previous concepts of ambidexterity. Future qualitative studies could integrate external organizational stakeholders to better understand the external dynamics of ambidextrous decision-making within AI deployment in organizations.

Second, the sample is predominantly German, restricting the generalizability of results to other cultural or regulatory contexts. AI-related decision dilemmas and its underlying ethical considerations may differ across regions and cultures due to varying norms, values, and governance structures. Comparative cross-cultural studies could deepen our understanding of how organizations respond ethically to these decision dilemmas with a more global perspective.

Third, while our analysis primarily draws on sequential and structural ambidexterity, future research could benefit from a deeper integration of additional ambidexterity concepts. In particular, the third decision dilemma – balancing standardization with personalized approaches – suggests that tensions between exploration and exploitation often arise within individual roles and tasks. Applying the concept of contextual ambidexterity, which emphasizes individuals' discretion in managing this balance within a given organizational context, could provide another perspective of how decision dilemmas can be navigated ethically at the micro level.

6. Conclusion

This work examines decision dilemmas that may arise when organizations integrate AI technologies into their operations and analyzes how they respond to such decision dilemmas in an ethical manner. Strategic choices regarding AI deployment often hinge on balancing exploration—experimenting with AI to generate novel business value—and exploitation—optimizing existing processes through AI. The study highlights that, in the context of AI, achieving this balance, known as ambidexterity, is closely intertwined with ethical considerations due to the sometimes unintended consequences of AI deployment for individuals and organizations. Moreover, this study contributes to the literature by outlining four central decision dilemmas that organization face when deploying AI and offers in-depth insights into how organizations navigate distinct decision dilemmas ethically. In doing so, this study provides ideas on how existing concepts of ambidexterity might be extended and enriches AI ethics dialogue in IS research with empirical insights.

Declaration of interest statements

Consent for publication

All authors have read and agreed to the final version of the manuscript.

Declarations of interest

None.

Funding

No funding was received.

References

- Alsheibani, S. A., Cheung, Y., Messom, D. C., & Alhosni, M. (2020). Winning AI Strategy: Six-Steps to Create Value from Artificial Intelligence. *Americas Conference on Information Systems (AMCIS)*.
- Ayoub, K., & Payne, K. (2016). Strategy in the Age of Artificial Intelligence. *Journal of Strategic Studies*, 39(5–6), 793–819. <https://doi.org/10.1080/01402390.2015.1088838>
- Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, 185(4), 725–740. <https://doi.org/10.1007/s10551-023-05339-7>
- Baptista, J., Stein, M.-K., Klein, S., Watson-Manheim, M. B., & Lee, J. (2020). Digital work and organisational transformation: Emergent Digital/Human work configurations in modern organisations. *The Journal of Strategic Information Systems*, 29(2), 101618. <https://doi.org/10.1016/j.jsis.2020.101618>
- Bareis, J., & Katzenbach, C. (2022). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, 47(5), 855–881. <https://doi.org/10.1177/01622439211030007>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). MANAGING ARTIFICIAL INTELLIGENCE. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75(1), 653–675. <https://doi.org/10.1146/annurev-psych-030123-113559>
- Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. *Sustainability*, 13(4), 1974. <https://doi.org/10.3390/su13041974>
- Brock, J. K.-U., & Von Wangenheim, F. (2019). Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. *California Management Review*, 61(4), 110–134. <https://doi.org/10.1177/1536504219865226>

Cao, L., Sarkar, S., Ramesh, B., Mohan, K., & Park, E. H. (2024). Shift of ambidexterity modes: An empirical investigation of the impact of artificial intelligence in customer service. *International Journal of Information Management*, 76, 102773. <https://doi.org/10.1016/j.ijinfomgt.2024.102773>

Chan, C. M. L., Teoh, S. Y., Yeow, A., & Pan, G. (2019). Agility in responding to disruptive digital innovation: Case study of an SME. *Information Systems Journal*, 29(2), 436–455. <https://doi.org/10.1111/isj.12215>

Chanas, S., Myers, M. D., & Hess, T. (2019). Digital transformation strategy making in pre-digital organizations: The case of a financial services provider. *The Journal of Strategic Information Systems*, 28(1), 17–33. <https://doi.org/10.1016/j.jsis.2018.11.003>

Chen, A., Wang, C., & Zhang, X. (2023). Reflection on the equitable attribution of responsibility for artificial intelligence-assisted diagnosis and treatment decisions. *Intelligent Medicine*, 3(2), 139–143. <https://doi.org/10.1016/j.imed.2022.04.002>

Chen, Y., Liu, H., & Chen, M. (2022). Achieving novelty and efficiency in business model design: Striking a balance between IT exploration and exploitation. *Information & Management*, 59(3), 103268. <https://doi.org/10.1016/j.im.2020.103268>

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60, 102383. <https://doi.org/10.1016/j.ijinfomgt.2021.102383>

Constantiou, I., Mayur, J., & Stelmaszak, M. (2023). Organizations as Digital Enactment Systems: A Theory of Replacement of Humans by Digital Technologies in Organizational Scanning, Interpretation, and Learning. *Journal of the Association for Information Systems*, 24(6), 1170–1798.

David, F. R. (2011). *Strategic management: Concepts and cases* (Third edition). Prentice Hall; Pearson, Upper Saddle River.

De Visser, M., De Weerd-Nederhof, P., Faems, D., Song, M., Van Looy, B., & Visscher, K. (2010). Structural ambidexterity in NPD processes: A firm-level assessment of the impact of differentiated structures on innovation performance. *Technovation*, 30(5–6), 291–299. <https://doi.org/10.1016/j.technovation.2009.09.008>

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>

Eldredge, N. (1972). Punctuated Equilibria: An Alternative to Phyletic Gradualism. *Models in Paleobiology/Freeman*.

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*, 24(5), 1709–1734. <https://doi.org/10.1007/s10796-021-10186-w>

Floridi, L., & Cows, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.8cd550d1>

Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>

Gioia, D. (2021). A Systematic Methodology for Doing Qualitative Research. *The Journal of Applied Behavioral Science*, 57(1), 20–29. <https://doi.org/10.1177/0021886320982715>

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, 16(1), 15–31. <https://doi.org/10.1177/1094428112452151>

Goossen, M. C., Bazzazian, N., & Phelps, C. (2012). Consistently Capricious: The Performance Effects of Simultaneous and Sequential Ambidexterity. *Academy of Management Proceedings*, 2012(1), 16311. <https://doi.org/10.5465/AMBPP.2012.16311abstract>

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, 29(2), 101614. <https://doi.org/10.1016/j.jsis.2020.101614>

Gupta, A. K., Smith, K. G., & Shalley, C. E. (2006). The Interplay Between Exploration and Exploitation. *Academy of Management Journal*, 49(4), 693–706. <https://doi.org/10.5465/amj.2006.22083026>

Hafner, A., Hammerschmidt, T., & Stolz, K. (2024). Analyzing The Use of Ethical Theories Within AI Ethics Research: A Systematic Scoping Review. *Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI)*. Proceedings of the 19th Internationale Tagung Wirtschaftsinformatik (WI), Wuerzburg, Germany.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>

Hammerschmidt, T., Hafner, A., Stolz, K., Passlack, N., Posegga, O., & Gerholz, K.-H. (2025). A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-025-10596-0>

Hammerschmidt, T., Stolz, K., & Posegga, O. (2024). How Leaders' Ambidexterity and Literacy Impact the AI Capabilities of Organizations. *Proceedings of the 32nd European Conference on Information Systems (ECIS)*. https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/14/

Hawkins, W., & Mittelstadt, B. (2023). The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 261–270. <https://doi.org/10.1145/3593013.3593995>

Hess, T., Matt, C., Benlian, A., & Wiesböck, F. (2020). Options for Formulating a Digital Transformation Strategy. In R. D. Galliers, D. E. Leidner, & B. Simeonova (Eds.), *Strategic Information Management* (5th ed., pp. 151–173). Routledge. <https://doi.org/10.4324/9780429286797-7>

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3), 101772. <https://doi.org/10.1016/j.jsis.2023.101772>

Jansen, J. J. P., George, G., Van Den Bosch, F. A. J., & Volberda, H. W. (2008). Senior Team Attributes and Organizational Ambidexterity: The Moderating Role of Transformational Leadership. *Journal of Management Studies*, 45(5), 982–1007. <https://doi.org/10.1111/j.1467-6486.2008.00775.x>

Jöhnk, J., Weißert, M., & Wyrski, K. (2021). Ready or Not, AI Comes—An Interview Study of Organizational AI Readiness Factors. *Business & Information Systems Engineering*, 63(1), 5–20. <https://doi.org/10.1007/s12599-020-00676-7>

Kegel, M. M., & Ghanem, L. (2024). You Did That on Purpose! An Investigation of the Knobe Effect in Human-Robot Interactions. *Proceedings of the 57th Hawaii International Conference on System Sciences*.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>

Ling, Y., Hammond, M., & Wei, L.-Q. (2022). Ethical leadership and ambidexterity in young firms: Examining the CEO-TMT Interface. *International Entrepreneurship and Management Journal*, 18(1), 25–48. <https://doi.org/10.1007/s11365-020-00695-6>

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>

Marabelli, M., & Galliers, R. D. (2017). A reflection on information systems strategizing: The role of power and everyday practices. *Information Systems Journal*, 27(3), 347–366. <https://doi.org/10.1111/isj.12110>

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1), 71–87.

Mayer, A.-S., Strich, F., University of Passau (Germany), Fiedler, M., & University of Passau (Germany). (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, 239–257. <https://doi.org/10.17705/2msqe.00036>

Milanez, A. (2024). *The role of bargaining and discrimination in the gender wage gap in France: A cross-country perspective* (OECD Social, Employment and Migration Working Papers 315; OECD Social, Employment and Migration Working Papers, Vol. 315). <https://doi.org/10.1787/1fd68687-en>

Mirbabaie, M., Brendel, A. B., Faculty of Business and Economics, Technische Universität Dresden, Hofeditz, L., & Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1), 726–753. <https://doi.org/10.17705/1CAIS.05034>

Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., & Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets*, 32(1), 73–99. <https://doi.org/10.1007/s12525-021-00496-x>

Mom, T. J. M., Chang, Y.-Y., Cholakova, M., & Jansen, J. J. P. (2019). A Multilevel Integrated Framework of Firm HR Practices, Individual Ambidexterity, and Organizational Ambidexterity. *Journal of Management*, 45(7), 3009–3034. <https://doi.org/10.1177/0149206318776775>

Morton, J., Amrollahi, A., & Wilson, A. D. (2022). Digital strategizing: An assessing review, definition, and research agenda. *The Journal of Strategic Information Systems*, 31(2), 101720. <https://doi.org/10.1016/j.jsis.2022.101720>

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, 17(1), 2–26. <https://doi.org/10.1016/j.infoandorg.2006.11.001>

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 64(5), 904–938. <https://doi.org/10.1177/0018720820960865>

O'Reilly, C. A., & Tushman, M. L. (2004). The ambidextrous organization. *Harvard Business Review*, 82(4), 74–83.

O'Reilly, C. A., & Tushman, M. L. (2011). Organizational Ambidexterity in Action: How Managers Explore and Exploit. *California Management Review*, 53(4), 5–22. <https://doi.org/10.1525/cmr.2011.53.4.5>

Pan, S.-L., Nishant, R., Tuunanen, T., & Choudrie, J. (2024). Is AI a strategic IS? Reflections and opportunities for research. *The Journal of Strategic Information Systems*, 33(4), 101866. <https://doi.org/10.1016/j.jsis.2024.101866>

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022). Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning

Stakeholder-Centered Solutions. *CHI Conference on Human Factors in Computing Systems*, 1–22.
<https://doi.org/10.1145/3491102.3517672>

Passlack, N., Hammerschmidt, T., & Posegga, O. (2024). It was(n't) me: Vignette experiments on managers' responsibility attribution in AI-advised decision-making. *Behaviour & Information Technology*, 1–30. <https://doi.org/10.1080/0144929X.2024.2431050>

Passlack, N., Heyder, T., Klemm, F., & Posegga, O. (2023). How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success. *Forty-Fourth International Conference on Information Systems*. Forty-Fourth International Conference on Information Systems, Hyderabad, India.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot* (arXiv:2302.06590). arXiv.
<http://arxiv.org/abs/2302.06590>

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, 43(1), iii–ix.

Raisch, S., & Birkinshaw, J. (2008). Organizational Ambidexterity: Antecedents, Outcomes, and Moderators. *Journal of Management*, 34(3), 375–409. <https://doi.org/10.1177/0149206308316058>

Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, 33(1), 100925. <https://doi.org/10.1016/j.hrmr.2022.100925>

Romanelli, E., & Tushman, M. L. (1994). ORGANIZATIONAL TRANSFORMATION AS PUNCTUATED EQUILIBRIUM: AN EMPIRICAL TEST. *Academy of Management Journal*, 37(5), 1141–1166. <https://doi.org/10.2307/256669>

Rosing, K., Frese, M., & Bausch, A. (2011). Explaining the heterogeneity of the leadership-innovation relationship: Ambidextrous leadership. *The Leadership Quarterly*, 22(5), 956–974.
<https://doi.org/10.1016/j.leaqua.2011.07.014>

Sarker, S., Xiao, X., Copenhagen Business School, Beaulieu, T., Utah State University, Lee, A. S., & Virginia Commonwealth University. (2018). Learning from First-Generation Qualitative Approaches in the IS Discipline: An Evolutionary View and Some Implications for Authors and Evaluators (PART 1/2). *Journal of the Association for Information Systems*, 19, 752–774. <https://doi.org/10.17705/1jais.00508>

Shahid, I., Anwar Ul Haq, M., Di Virgilio, F., & Ahmed, M. A. (2023). Ethical leadership and individual ambidexterity: The mediating role of equity sensitivity and the moderating role of openness to experience. *Management Research Review*, 46(12), 1727–1747. <https://doi.org/10.1108/MRR-06-2022-0418>

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22(5), 648–663. <https://doi.org/10.1080/1369118X.2019.1568515>

Shankar, V. (2024). Managing the Twin Faces of AI: A Commentary on “Is AI Changing the World for Better or Worse?” *Journal of Macromarketing*, 44(4), 892–899. <https://doi.org/10.1177/02761467241286483>

Smith, W. K., & Tushman, M. L. (2005). Managing Strategic Contradictions: A Top Management Model for Managing Innovation Streams. *Organization Science*, 16(5), 522–536. <https://doi.org/10.1287/orsc.1050.0134>

Tai, J. C. F., Wang, E. T. G., & Yeh, H.-Y. (2019). A study of IS assets, IS ambidexterity, and IS alignment: The dynamic managerial capability perspective. *Information & Management*, 56(1), 55–69. <https://doi.org/10.1016/j.im.2018.07.001>

Van De Wetering, R., Mikalef, P., & Dennehy, D. (2022). Correction to: Artificial Intelligence Ambidexterity, Adaptive Transformation Capability, and Their Impact on Performance Under Tumultuous Times. In S. Papagiannidis, E. Alamanos, S. Gupta, Y. K. Dwivedi, M. Mäntymäki, & I.

O. Pappas (Eds.), *The Role of Digital Technologies in Shaping the Post-Pandemic World* (Vol. 13454, pp. C1–C1). Springer International Publishing. https://doi.org/10.1007/978-3-031-15342-6_38

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4), 97–112. <https://doi.org/10.9785/cr-2021-220402>

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>

Wang, N., Hu, H., Liang, H., & Xue, Y. (2024). How to Develop IT Ambidexterity: Insights from the Perspective of CIO-TMT Mutual Communication. *Journal of the Association for Information Systems (JAIS)*, 164.

Watson, G. J., Desouza, K. C., Ribiere, V. M., & Lindič, J. (2021). Will AI ever sit at the C-suite table? The future of senior leadership. *Business Horizons*, 64(4), 465–474. <https://doi.org/10.1016/j.bushor.2021.02.011>

Wimelius, H., Mathiassen, L., Holmström, J., & Keil, M. (2021). A paradoxical perspective on technology renewal in digital transformation. *Information Systems Journal*, 31(1), 198–225. <https://doi.org/10.1111/isj.12307>

Wirtz, J. (2020). Organizational Ambidexterity: Cost-Effective Service Excellence, Service Robots, and Artificial Intelligence. *Organizational Dynamics*, 49(3), 100719. <https://doi.org/10.1016/j.orgdyn.2019.04.005>

Yang, J., Amrollahi, A., & Marrone, M. (2024). Harnessing the Potential of Artificial Intelligence: Affordances, Constraints, and Strategic Implications for Professional Services. *The Journal of Strategic Information Systems*, 33(4), 101864. <https://doi.org/10.1016/j.jsis.2024.101864>

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101, 104327. <https://doi.org/10.1016/j.jesp.2022.104327>

Appendices

Appendix A: Qualitative Questionnaire

Table A.1 provides an overview of our semi-structure interview guide.

Section	Questions	Time
Introduction	Welcome, Introduction, Technical Setup, Agreement for Recording	3 minutes
First Section (Working culture, business nature, and task and decision responsibilities)	- How would you describe the working and business nature of the company and industrial sector you are currently working in? - How would you describe the culture in your direct working environment (team-level) – especially with regard to the openness toward novel technologies? Does the culture between your team and other parts of the company vary, and why? - How would you describe your primary responsibilities for activities, technologies, and other people in your current positions? - Please describe the decisions that you have to make as part of your job. Would you rate those decisions as having a strategic impact on the company or other organizational stakeholders, and why?	8 minutes
Second Section (AI understanding, scenarios, and dilemmas)	- How would you describe AI to a friend or colleague? - What do you think are the unique features of human beings in comparison to AI? - Please describe (at least, one) examples where AI technologies are discussed, developed, used, trained, or implemented in your working context? [<i>Explanation: These can be AI technologies with which you interact directly as a user or with which others (intend to) interact with your indirect involvement (e.g., as supervisor or trainer)</i>]. - Please describe the specific tasks and goal of AI usage. - What strategic impact to you expect from the AI usage? - Please describe whether the AI technologies are controlled by the organization (i.e., initiated by the subordinate) and how or are used without supervision (uncontrolled usage) and why.	10 minutes
Third Section (Managerial strategies and underlying measures and requirements)	- What mechanism (processes, guidelines, strategies) does your organization have to guide with the AI examples you mentioned? - What do you think hinders an organizations and workforce’s development of abilities needed for realizing strategic goals with AI deployment and usage? - Please describe the ideal AI strategy for an organization: What can organizations do to benefit from AI strategically and to reduce the negative consequences of AI usage?	7 minutes
Fourth Section (Decision dilemmas illustrated through four effects of ethically managing human-AI interaction)	<i>Explanation: In the following, we would like to discuss the following four effects based on the work of Heyder et al. (2023) that can occur when ethically managing humans and AI in organizations as they illustrate decision dilemmas:</i> <i>Responsibility Attribution Effect</i> which implies that humans transfer their responsibility for decisions and actions (totally or partly) to AI technologies deontology in situations where they do not fully understand AI’s outputs, cannot control AI’s actions, or feel unpleasant with the AI’s outcomes, wherefore they curtail their own ethical being/value (<i>virtue ethics</i>). <i>Control Coherence Effect</i> which implies that humans act in non-alignment with existing principles and norms in situations (non-compliance with <i>deontology</i>) by for instance using loopholes) where AI overtakes leadership functions, wherefore humans try to compensate a perceived imbalance of control and autonomy of the AI compared to themselves (protect <i>virtue ethics</i>). <i>Altruistic Behavior Effect</i> which implies that humans use their ethical knowledge based on their virtue ethics to question decisions made by the AI’s (<i>deontology</i>) to achieve greater perceived fairness for others being affected by AI’s outcomes (<i>consequentialism</i>). <i>Behavior Adoption Effect</i> which implies that humans adopt their behavior based on the interaction with the AI (and their underlying ethical principles following <i>deontology</i>) that impacts their moral values and <i>virtue ethics</i>, which can be good or bad. <i>For each effect, we ask the following three questions:</i> - Why there are or could be a similar effect in your working environment? Please describe a situation, in which this effect might occur. - Who should be responsible to deal with this effect (e.g., by making certain decisions?) - What are requirements to deal with this effect appropriately?	15 minutes
Outro	Next Steps, Open Questions, Feedback, Acknowledgements	2 minutes

Table A.1: Semi-structured Interview Guide

Appendix B: Overview of Dilemmas Related to AI Tasks and Contexts

Table B.1 summarizes the different decision dilemmas mentioned above dependent on the different types of tasks of the AI deployment context (see Appendix C for details).

Task & Context	Strategy	Decision Dilemma	Ethical Considerations
Decision-making tasks such as pre-selection of candidates with AI; candidate assessment with AI in recruiting; and AI bots advising within cybersecurity.	Organizations implement AI in small projects to prove its potential for digitalizing existing processes of data aggregation and analysis, previously done by humans (exploitation), with the final decision remaining with humans. This strategy can be driven by external pressures of competitors using AI for gaining strategic advantages, by being more cost-effective, for instance (punctuated equilibrium).	Main dilemmas arise from building trust in AI to foster AI adoption while similarly aiming to avoid overtrust in AI, thus ensuring that users do not blindly follow AI recommendations linked to overreliance or critical gaps in human judgment. Another main dilemma concerns juxtaposing leveraging AI's capabilities with the potential for ethical violations .	- Ensure AI algorithms used provide fair and unbiased evaluations; maintain an objective and standardized approach while avoiding dehumanizing the process (e.g., by advocating for state-level legal standards and stronger accountability in AI development). - Ensure that users understand the role of AI in the process; address concerns of opacity in AI decision-making, particularly in high-stake fields such as cybersecurity (e.g., through training to realistically assess AI's abilities).
Creation and information tasks such as content creation; information search; and programming with generative AI (e.g., ChatGPT, Dall-E, CoPilot)	AI use is often unsupervised, with individual employees improving performance (exploitation), influenced by media hype that can create sudden surges of attention and investment and disrupting the status quo (punctuated equilibrium).	Main dilemmas concern balancing the need for personalized approaches with the standardization inherent in generative AI tools. Another main dilemma arises around shared responsibility versus clear allocation of individual responsibility when outputs contain errors or infringe intellectual property rights.	- Protect sensitive data to prevent it from being unintentionally exposed in training or use (e.g., by providing training on how current law impacts generative AI use, such as the European AI Act). - Clearly communicate the use of generative AI tools to clients and internal stakeholders (e.g., by establishing guidelines for generative AI use), especially in contexts where outputs may influence decision-making.
Prediction tasks such as algorithmic prediction platforms and AI simulations; anomaly detection algorithms; and fraud detection in ticket sales.	AI is used to drive innovation and explore novel areas for business and match solutions (exploration).	Main dilemmas concern balancing that AI predictions align with standardization while safeguarding personalized approaches , especially when suggestions are followed uncritically. Another decision dilemma concerns leveraging AI's capabilities for effective use over ethical violations , meaning that while AI has great potential for improving processes in different ways, there is a risk it may breach privacy or other ethical boundaries.	- The global context and the autonomous and often non-transparent actions of different AI prediction tools makes it difficult to identify possible issues and creates further challenges to pinpointing responsibilities when issues or unintended consequences arise. - Ethical concerns arise regarding data ownership, control, and potential misuse. Thereby, clear contracts and safety protocols may be helpful.
Optimization tasks such as predictive maintenance; autonomous manufacturing; AI drone operations; and damage assessment with AI.	AI deployment is driven by concerns of disruption (punctuated equilibrium); making organizations enhance their operational efficiency to improve quality and reduce costs (exploitation).	Main dilemmas center on balancing standardization to increase efficiency (e.g., cost-saving goals) with personalized approaches related to job security and maintaining one's role identity.	- Balance the efficiency gains from AI with environmental responsibility and resource sustainability (e.g., by invest in critical AI assessment processes and quality controls). - Ensure operational improvements (e.g., increased efficiency) do not harm employees (e.g., their job satisfaction and well-being). - Ensure that AI-driven assessments do not undermine employee expertise or quality.

Physical tasks, such as healthcare robots providing support in elderly care or in assisting operations, or robots used in manufacturing environments for material handling.	AI is used to explore how AI-based robots can assist and enhance physical tasks (exploration).	Main dilemmas arise from determining accountability in scenarios where robots act autonomously, balancing shared responsibility with clear allocation of individual responsibility, particularly when (unintended) negative consequences appear and when AI is deployed in high-stakes environments like healthcare.	- There is a risk of anthropomorphizing robots to a degree that may affect individual's decision-making and ethical considerations in interacting with them as individuals may perceive AI not solely as technology but also attributing human characteristics toward AI. - Get users on board when developing robots that carry out physical tasks to address ethical concerns early in the development process and to foster learning-by-doing. - AI robots may enhance human's natural abilities with questions on who is accountable when there are unintended outcomes.
Communicative and supportive tasks such as chatbots for customer support or learning agents; and matching algorithms for customer preferences.	Organizations are digitalizing existing processes with AI technologies (exploitation).	Main dilemmas center on leveraging AI's capabilities for effective use to support certain tasks or services, with the risk of undermining human connection or moral values (creating ethical violation). In additionally, in matching algorithms, dilemmas arise when prioritizing efficiency through standardization over individual preferences (personalized approaches).	- AI should be used only in ways that ensure its outputs are applied responsibly (e.g., by providing clear explanations of AI capabilities and limitations to users). - Responsibility for AI-related outcomes should rest with those who develop and provide the tools, ensuring that developers and AI providers take responsibility and communicating this to customers.

Table B.1 Overview of Dilemmas Related to AI Tasks and Contexts

Appendix C: Detailed Description of Use Cases, Decision Dilemmas, and Measures per Interview

Company	DEDUCTIVE FRAMING OF THE ORGANIZATION			INDUCTIVE CODING OF THE USE CASES & DILEMMAS OF AI DEPLOYMENT			
	Organizational Factors	Technological Factors	Environmental Factors	AI Use Cases	AI strategy	Dilemmas	Measures
L1 (COO, AI Software Start-up)	• Strategic Alignment: AI tools should only be implemented by customers when relevant data is available. The focus should be on data quality and availability, rather than on the AI model itself. • Knowledge: Analytical thinking and technical expertise are crucial for understanding and leveraging AI tools effectively. • Culture: Customers should foster a supportive culture and have a clear understanding of how AI functions and contributes to solutions.	• Resources: Organizations need access to data that is not only relevant to the AI use case but also structured and of high quality to ensure effective model training. • Data: Providers must invest significant time in preparing high-quality training data, as this is often the most time-consuming part of building AI models (accounting for approximately 90% of the effort). Many organizations lack a clear data structure, making it difficult for data owners to gain insights or control over the data.	• Non-Discrimination-Principle: Data quality is more critical than the AI model itself, as poor-quality data can lead to ineffective or biased outcomes. • Fairness-Principle: Fairness in AI is influenced by cultural context, meaning that AI models developed in Europe may not be fair or suitable for regions like South America due to differing values and norms.	#1 Algorithmic-based Predictions in Marketing: Evaluating the effectiveness of product images by assessing how well they resonate with customers and the emotions they evoke. #2 AI tools for picture generation and Photoshop: A fictitious example for creating and manipulating images, such as enhancing product photos or generating visual content for marketing purposes.	The organization focuses on exploitation strategies by moving beyond traditional SAP systems, aiming to digitally transform towards e-commerce, digital experience platforms, and chatbots, while exploring new product offerings to drive innovation and enhance customer engagement.	#1 Shared responsibility vs. clear allocation of individual responsibility: The concept of sense of honor versus searching for excuses pertains to individual accountability and responsibility, a fundamental aspect of education that remains relevant regardless of AI usage. #2 Effective use of AI's capabilities vs. ethical violation: One can enhance a CV picture using Photoshop for professional purposes, like improving clarity or adjusting lighting. or use Photoshop to fabricate fake news or mislead others through manipulated imagery.	• Customer Data Preparation: Teach customers how to prepare their own data instead of doing it for them, which also supports data security practices. • Data Governance: Establish clear data governance principles, ensuring proper data ownership and control, which are essential for AI implementation in organizations. • Fairness in AI: Define clear rules and guidelines for AI operations to ensure fairness, emphasizing that AI itself does not discriminate, but its application can if not properly managed. • Strategic Decision-Making: Take the necessary time to make informed strategic decisions and avoid rushing into implementations that could lead to failure. • Ethical Awareness: Promote moral values and raise awareness on ethics to prevent the exploitation of AI loopholes or unethical usage.
L2 (CPIO, Automotive Supplier)	• Strategic alignment: The organization maintains a cautious culture, with an emphasis on cost-efficiency in production processes while adopting a reflective and deliberate approach to AI integration. • Knowledge: Continuous knowledge development and skills	• Data: Data resources are predominantly geared toward automotive production processes, with limited integration in HR functions (lack of an overarching data architecture). • IT Resources: Advanced infrastructure exists for automating	• Regulation: Compliance with social media guidelines, IT security policies, and HR-related legal frameworks. • Security-Principle: Ensuring individuals' personal data is handled responsibly.	#1 AI recruiting agent: So far, conceptual work on how AI can assist the selection process, particularly for differentiating between blue-collar workers (where skills are more easily quantifiable) and white-collar workers (where more subjective, empathetic	In the context of autonomous manufacturing, focus on exploitation to maximize cost-efficiency in core production processes (ensuring competitive advantages). When it is about secondary, non-	#1 Shared responsibility vs. clear allocation of individual responsibility: Tensions arise when recruiters must reject well-known or referred candidates based on AI recommendations, leading to interpersonal conflicts and ethical considerations.	• Online Learning: Introduce online learning tools and platforms to encourage continuous professional development and upskilling. • Foster Dialogue: Engage employees in personal conversations to understand and address their concerns about job security and AI replacement. • Organizational Structures: Implement a matrix organizational structure to normalize role changes and foster flexibility, ensuring

	enhancement are integral to the company's long-term vision. • Culture: A strong focus on fostering learning, personal growth, and professional development among employees.	manufacturing, but HR processes lack comparable technological investment.	• Fairness-Principle: Promoting equitable decision-making processes in recruitment.	attributes are essential). #2 Autonomous manufacturing: AI-driven software integrated into manufacturing to enhance efficiency, reduce waste, and minimize costs (as competitive advantage).	core business processes, such as recruiting, the organization aims to experiment with novel technologies to gain novel insights and increasing learning with the technology (exploration) without compromising human oversights.	#2 Standardization vs. personalized approaches: Shifting roles and responsibilities as employees adapt to increased AI automation, creating challenges in balancing standardized efficiency and personalized decision-making.	employees view technology as an enabler rather than a threat.
L3 (CEO, Hiring Agency)	• Strategic alignment: No immediate sense of urgency to integrate AI into the organization. • Knowledge: Minimal AI-specific expertise, though there is an awareness of its potential to enhance recruitment processes. • Culture: Employees harbor misunderstandings about AI, perceiving it as a threat to job security, hindering acceptance.	• Data: Recruitment decisions currently rely on manual methods, such as reviewing CVs and LinkedIn profiles to match job requirements. • IT Resources: Lack of allocated resources for AI development or implementation (generally not much IT-related resources).	• Beneficence-Principle: Addressing the potential depersonalization of the recruitment process due to excessive reliance on AI.	AI recruiting agent: The proposed AI agent will conduct initial interview assessments to evaluate candidates' fit. Successful candidates will progress to a second interview with human recruiters to provide a more personalized evaluation.	The organization focuses on exploitation strategies, planning to optimize existing recruitment processes using AI to improve efficiency and streamline workflows.	Standardization vs. personalized approaches: Striking a balance between AI automation (e.g., for conducting personality assessments) and maintaining human involvement (e.g., for empathy and intuitive decision-making).	• Enhance Trust: Foster a culture of trust in AI by providing employees with hands-on experiences and training to familiarize them with AI tools. • Enhance Transparency: Promote transparency about how AI is used in recruitment to alleviate concerns among candidates and staff. • AI Guidelines: Establish clear legal and ethical guidelines for using AI in recruitment to ensure compliance and maintain fairness.
L4 (CEO, Ad & Media Agency)	• Strategic alignment: Strong support from top management with high customer expectations driving AI adoption.	• Data: Lack of structured, data-driven processes limits certain (non-generative) AI applications.	• Guidelines & Regulation: GDPR, Marketer Code of Conduct • Transparency-Principle:	Content creation with ChatGPT and Dall-E: Generative AI tools are employed to create text and visual content for clients. AI is	The organization focus on punctuated equilibrium, utilizing generative AI	#1 Standardization vs. personalized approaches: The need to produce unique, high-value content to differentiate from what clients could	• Customer Experience: Develop a strong customer-centric approach to create bespoke, high-quality experiences that differentiate the agency from generative AI-only

	• Knowledge: Increasing need for expertise in prompt engineering to fully leverage generative AI tools. • Culture: Agile culture supports adaptability and innovation (typical for this type of industry).	• IT Resources: No specific internal IT infrastructure needed to capitalize on AI capabilities.	Determine how openly the agency should disclose its use of generative AI tools to clients. • Beneficence-Principle: Balancing growth with employee welfare (organizational efficiency against the impact on employees whose roles are replaced).	positioned as a competitive advantage for enhancing creativity and efficiency.	tools to optimize workflows and content as the current hype about ChatGPT is stressing the agencies daily work.	create using generative AI tools themselves. Termination of existing copywriters as AI replaces their roles, leading to workforce challenges. #2 Shared responsibility vs. clear allocation of individual responsibility: If content produced using generative AI is successful, who receives credit: the AI tools, the employee prompting the AI, or the agency? If content fails to meet expectations, how is accountability distributed?	solutions (customer feel-good management) • Skill-Strategies: Implement workforce reallocation strategies, including reskilling or transitioning employees affected by AI integration. • Transparent Communication: Foster transparency by clearly communicating the role of AI in content creation to clients. • AI Guidelines: Establish clear guidelines for attributing success or failure when using AI-generated content. • Style guide: Set guidelines for adapting AI-generated content to the company's style and tone.
L5 (CHRO, Insurance Company)	• Strategic alignment: High level of top management support and strong alignment between AI processes and strategic objectives. However, there is limited integration of AI products into existing workflows. • Knowledge: The organization demonstrates agile competencies and a high level of digital literacy among its workforce. • Culture: Open mindset toward innovation,	• Data: High availability of data, with ongoing progress toward establishing a centralized data management system. • IT Resources: Insufficient IT capacity to scale and integrate AI pilot projects into operational workflows effectively.	• Regulation: HR law, Car liability insurance • Transparency-principle: Transparency in AI decision-making is critical to avoid discrimination or unfair outcomes. • Fairness-Principle: Quality of assessments and outcomes need to be continuously monitored. • Beneficence-Principle:	#1 Damage assessment with AI: Pilot project in the claims management department utilizing an AI-driven image detection tool to estimate damage. #2 AI recruiting agent: Testing an AI bot for candidate selection, including an A/B/C classification of candidates.	The organization focuses on exploitation strategies, leveraging AI primarily to improve and optimize existing core processes (esp. damage regulation) but also to improve processes besides the core value chain (esp. recruiting).	#1 Standardization vs. personalized approaches: Conflicts may arise between cost-saving measures and enhanced assessment quality versus employees' evolving roles and responsibilities. #2 Effective use of AI's capabilities vs. ethical violation: Dependence on external AI providers presents risks, including black-box concerns and potential biases in AI systems (difficult to assess internally).	• Agile Culture: Foster an agile workforce by implementing continuous learning and reskilling programs. • Flat Hierarchies: Empower employees and promote flatter organizational hierarchies. • Open Culture: Cultivate an open-minded, innovation-friendly culture. • Public Perceptions: Enhance public communication to build trust and transparency. • AI Knowledge: Invest in internal knowledge development to critically evaluate AI processes.

	though tempered by a cautious industrial culture.		Employees may feel devalued as their roles evolve due to AI integration (also customers could perceive the process as less empathetic).			Reliance on external innovation brings challenges in maintaining transparency and fairness.	
L6 (CMO, Renewable Energy Provider)	• Strategic alignment: Strong top management support with partial integration of AI into workflows. There is a lot of unsupervised use of AI technologies. • Knowledge: High intrinsic motivation among employees to learn and develop skills, supported by upskilling initiatives. • Culture: Employees are accustomed to change management practices, fostering adaptability.	• Data: Uneven availability and accessibility of relevant data (no central strategies for data management). • IT Resources: The organization prioritizes hiring for adaptability and personality traits over technical expertise, ensuring staff openness to change.	• Beneficence-Principle: Ethical implications of prioritizing efficiency over job satisfaction. • Fairness-Principle: Quality of assessments need to be continuously monitored. • Non-maleficence-Principle: Compliance with existing law must be ensured (e.g., European AI Act: labeling of contributions created by generative AI)	#1 AI drone operations: Detecting damages in solar fields, significantly increasing employee productivity (e.g., one employee now oversees multiple solar parks). #2 Content creation with ChatGPT: Leveraging ChatGPT for generating marketing campaigns, social media posts, and newsletters.	The organization emphasize exploitation strategies using AI to reduce time-consuming control processes when it is about their core processes as an energy provider, but experiments (rather unsupervised) with generative AI tools (exploration strategies) when its about their marketing.	#1 Standardization vs. personalized approaches: While AI tools improve organizational performance, employees may face role conflicts, such as transitioning from manual inspections to overseeing drone operations. #2 Standardization vs. personalized approaches: AI-generated content often lacks personalization and requires adaptation to align with the company's branding and emotional appeal. Over-reliance on AI might reduce creativity and critical thinking among employees.	• AI Contracts: Establish clear contractual agreements for AI implementation and employee roles. • AI Training: Develop employees' skills to take on novel tasks with training and skill initiatives and potential incentives. • Role Description: Define clear job roles and responsibilities as tasks evolve. • Role Models: Promote leadership that emphasizes personal development and responsibility.
L7 (CEO, Engineering Company)	• Strategic alignment: Strong fit between AI processes and organizational strategy, although customer readiness for AI	• Data: Machines are equipped with sensors for data tracking, but complexities and uncertainties in data ownership hinder further progress.	• Regulation: Struggle with different regulations as a multinational company (specifically	AI detection in manufacturing: Predicting machine wear and tear to prevent breakdowns and reduce production	The organizations focus on exploitation as they fear to be replaced by disruptive	Effective use of AI's capabilities vs. ethical violation: Dependence on external AI platforms could expose the organization to risks such as data misuse	• Legal Requirements: Advocate for state-level legal requirements to establish secure standards protecting against intellectual property theft and industry disruption.

	adoption remains a barrier. • Knowledge: Significant gaps in digital literacy, with ongoing efforts to recruit digitally skilled employees. • Culture: Lack of trust in the potential of AI technologies for the own organization, but an open culture of accepting failure facilitates initial experimentation with AI.	• IT Resources: Limited IT infrastructure and skilled personnel, leading to reliance on external IT providers.	Chinese vs. European law). • Security-Principle: Ethical concerns about data ownership and control. • Non-maleficence-Principle: Fear of replacement through disruptive AI players.	downtimes (predictive maintenance).	players from other markets (especially China), as their regulation is different to the European ones, making it easier to stay competitive (punctuated equilibrium).	or industry disruption by low-cost competitors.	• Data Contracts: Develop secure data-sharing agreements with external providers.
L8 (CEO, Pharmaceuticals)	• Strategic alignment: Within the organization there is limited management support for data analysis due to skepticism about AI and a preference for a non-profit-driven culture. • Knowledge: Data knowledge is limited, and there is resistance to tracking, primarily due to concerns about data security. • Culture: The organization fosters a respectful, communicative, and transparent culture with low hierarchies, emphasizing sustainability over profitability.	• Data: The organization resists tracking employee performance, focusing instead on open communication and acceptance, without monetary incentives. • IT resources: The organization's IT infrastructure is underdeveloped, largely due to management's reluctance toward AI adoption.	• Data security Principle: In a hypothetical scenario, leadership should not have the right to access personal data for controlling employees, as this could violate privacy rights and undermine trust within the organization.	#1 Data analytics for performance evaluation: The manager used data analytics to evaluate the performance and profit from certain products. #2 Digital skiing teacher: Conceptual thoughts on a tool with sensors in skiing shoes analyzes a skier's performance and provides real-time recommendations through earplugs, trained using insights from professional skiing experts.	The organization focuses on exploration strategies by designing innovative technologies like the digital skiing teacher, aiming to experiment with new concepts and push the boundaries of what is possible in the field.	#1 Standardization vs. personalized approaches: The manager used data analysis to evaluate performance but hide the use of AI, attributing decisions to gut feeling, to fit the organizational culture. #2 Shared responsibility vs. clear allocation of individual responsibility: In the case of a digital skiing teacher, AI may suggest actions like "give more power," which can lead to dangerous outcomes, as AI lacks awareness of the student's physical condition or safety.	• Motivation to Learn: Individuals should be motivated to understand and learn about algorithms to prevent overtrust in their outputs. • Transparency and Communication: To prevent the misuse of loopholes, make them transparent and communicate openly that they should not be exploited for the tool to function effectively. • Focus on Learning: Instead of just teaching specific skills, focus on teaching individuals how to learn and adapt in the context of AI.

L9 (CMO, News & Media Platform)	• Strategic Alignment: The organization tested ChatGPT but concluded there is no suitable AI process fit, using only basic algorithms for data analysis. There is a need for more support from the management level, and the organizational structure does not support seamless AI integration into workflows. The infrastructure and tools require strategic alignment for effective AI deployment. • Knowledge: The team possesses high technological skills, with expertise in psychology and mathematics, which is beneficial for AI-related tasks. • Culture: The organization is driven by innovation with an agile culture, and algorithms are not new; however, they tend to be traditional and do not see AI as relevant to their current use cases.	• Data: The organization faces a data ownership dilemma as they use a service on their website that collects data, but they do not fully understand the data flow or how the data is managed. • IT Resources: The current IT infrastructure is insufficient to implement clear AI processes, and the organization lacks skilled personnel and structured resources to support AI integration.	• Security-Principle: Organizations often face the problem of lacking ownership or rights over the data from their customers when using third-party services, which can lead to concerns about data security, privacy, and compliance with regulations.	#1 Content Creation in Journalism: Journalism uses ChatGPT to assist in generating content, such as articles or reports, to streamline the writing process. #2 Predictions in Marketing: Algorithms predict and suggest articles to publish in online magazines based on audience preferences and engagement data.	The organization focuses on an exploitation strategy by evaluating and testing AI algorithms to optimize their output, ensuring that existing processes are improved and efficiency is enhanced through the application of AI technologies.	#1 Standardization vs. personalized approaches: AI tools suggested promoting certain articles based on data-driven insights, but the organization preferred different parameters due to internal values and gut feeling. This illustrates the conflict between trusting data and relying on human judgment or organizational priorities. #2 Standardization vs. personalized approaches: Using AI for content creation may lead to homogenous outputs, reducing the uniqueness of the content. This risk arises as AI often generates content based on patterns, potentially limiting creativity and originality.	• Regular Meetings and Best Practice Sharing: Hold regular meetings to discuss potential AI use cases and share best practices, ensuring the organization stays up to date with emerging trends and technologies. • Pilot Testing: Conduct pilot testing to ensure AI implementation increases efficiency rather than creating inefficiencies through redundant checks or errors. • Governmental Regulations: Establish governmental regulations to address privacy concerns, as third-party providers may not prioritize data protection due to perceived disadvantages. • Reduce Hierarchy and Increase Diversity: Foster a more inclusive environment by reducing hierarchy and increasing diversity, which can drive successful AI transformation and innovation within the organization.
L10 (Leading Developer,	• Strategic alignment: The organization positions itself as a pioneer in AI adoption	• Data: Machines are equipped with sensors for real-time data tracking; however,	• Sustainability-Principle: Balancing technological	Smart home: Incorporating sensors into washing machines to	The organization focuses on exploration	Standardization vs. personalized approaches:	• AI Training: Promote transparency and user education about the benefits and limitations

Electronic Provider)	within the electronics industry, leveraging its innovative capabilities to remain at the forefront of technological advancements. • Knowledge: The workforce possesses extensive expertise in AI applications, complemented by a realistic understanding of the technology's limitations and potential challenges. • Culture: While the company fosters an open and innovative mindset, there is also a cautious approach to ensure responsible AI implementation.	progress is hindered by complexities in data ownership and associated legal uncertainties. • IT Resources: As a market leader, the company benefits from robust internal resources, including advanced technological infrastructure and strong political support for its initiatives.	advancements with sustainable resource utilization and environmental responsibility.	determine the appropriate amount of water usage based on the weight of the clothes, enhancing efficiency and sustainability	strategies, using AI to develop new capabilities and enhance product innovation.	Managing the risk of over-reliance on AI systems that could lead to skill erosion of individuals (particularly when its about the users of Smart home technologies), while simultaneously creating opportunities for upskilling in AI technologies (particularly for employees that are also users; dual role problematic).	of AI-enabled smart home technologies.
L11 (Project Lead, Training Institute)	• Strategic alignment: The organization's management aims to clearly explain the opportunities and benefits of, ensuring a shared understanding and gaining support across all levels of the organization. • Knowledge: Currently, the organization relies on external programmers but plans to train its personnel in the long term to improve efficiency and	• Resources: Initially, the organization relies on external providers to learn and gain expertise. However, it emphasizes the importance of hiring individuals with the necessary skills, as external personnel can slow down processes over time.	• Non-maleficence-Principle: AI should only be used in ways that ensure its output is applied responsibly, without causing harm to individuals or society.	#1 Learning Records Platform: A data-driven learning platform that supports personalized studying by tracking performance and adapting to individual learning progress. #2 Tracking of Pill Taking: A hypothetical system designed to remind elderly individuals to take their medication. The AI must be user-friendly, ensuring	The organization focuses on experimenting with new technologies and equipping individuals with AI knowledge through training and education (exploration strategy).	#1 Shared responsibility vs. clear allocation of individual responsibility: In a hypothetical scenario of controlling pill-taking, an AI system reminds patients to take their medication and tracks adherence. If a patient misses a dose and experiences adverse effects, medical personnel might shift blame to the AI system, questioning its reliability and absolving themselves of responsibility.	• User-Friendly Interface: AI systems should have an intuitive interface that enables users to learn about the tool through hands-on experience. • Engage Personnel and Managers: Sensitize both personnel and managers to the importance of AI, ensuring everyone is on board with its implementation. • Learn from External Providers: Start by learning from external providers, but gradually transition to in-house expertise for greater autonomy. • Ethics Board Oversight: Parameters should be

	streamline communication processes. • Culture: Embraces new technologies, works agilely, and stays consistently up-to-date with the latest advancements.			ease of use for all users, particularly those with limited technological experience.			established and monitored by an ethics board to ensure ethical compliance. • People Empowerment: Empowering individuals is crucial, but it should be based on their qualifications, allowing them to make decisions aligned with their own moral values.
L12 (Account Executive, Cloud Services)	• Strategic alignment: Specialized in strategic alignment • Knowledge: The team is highly tech-savvy, well-aware of AI's limitations, and actively engages in raising customer awareness of ethical considerations. Their responsibilities include leveraging explainable AI to clarify the technology's functions and impacts to clients. • Culture: The team demonstrates strong tech-affinity, coupled with curiosity and openness to exploring new technologies.	• Data: A lot of customer data available through cloud services • IT Infrastructure: High IT resources and security infrastructure as providing cloud services for data and model storages for a great number of customers.	• Non-discrimination-Principle: There is a critical issue that biases in underlying data models are reflected in algorithmic outputs, potentially perpetuating structural discrimination, such as associating minority groups with fraudulent behavior based on historically biased data; as they were controlled more often in the past.	Fraud detection in ticket sales: The use case involves utilizing historical ticket sales data, including instances of fraudulent transactions, to train a machine learning model. This model can then analyze new transactions in real-time, flagging potentially fraudulent purchases based on patterns identified in the training data.	An organization that provides consulting services to other organizations on their technology products primarily focuses on developing exploitation strategies to help their clients optimize processes and enhance productivity.	Standardization vs. personalized approaches: The output of the algorithm is typically a confusion matrix, which categorizes transactions based on probabilities, such as the likelihood of being fraudulent or legitimate. The challenge lies in balancing the thresholds to minimize false positives (blocking legitimate transactions) and false negatives (missing fraudulent ones), with decisions made on whether to flag or automatically block suspicious transactions.	• AI Training: Internal trainings on AI limits and the technologies ethical impacts to communicate it to customers • AI Contracts: Responsibilities must be contractually defined. Legally, it is unlikely to shift this responsibility to developers or those who train data models, and this is unlikely to change in the future.
L13 (AI Sales Lead, Technology Provider)	• Strategic alignment: Software innovation is at the core of the company's business value, with data recognized as a high-value strategic asset.	• Data: Relies on customers' data rather than owning proprietary datasets, focusing on data security and ethical usage.	• Regulation: Compliance with the European AI Act but facing gaps in industrial legal frameworks, particularly regarding accountability in	AI in coding (GitHub Copilot) and application development (OpenAI): Leveraging tools like GitHub Copilot and OpenAI Transformer Architecture to	Focus on exploration strategies as a central component of the business model, emphasizing constant	#1 Standardization vs. personalized approaches: Customers prioritize ease of use and intuitive AI models, whereas the organization must maintain critical knowledge of AI	• AI Training: Train employees to deeply understand and critically evaluate AI outcomes to ensure high-quality solutions. • System Design: Prioritize creating intuitive, user-friendly interfaces for customers while balancing

	• Knowledge: A highly skilled and digitally literate workforce supports cutting-edge developments. • Culture: Technological advancement is deeply ingrained in the company's culture, fostering high levels of tech affinity and openness to innovation.	• IT Resources: Adapting on open-source models rather than building novel ones. Extensive tools and platforms available for developers and data scientists, enabling efficient experimentation and application development.	multi-party contracts. • Accountability-Principle: Ensuring safe AI usage while maintaining responsibility for outcomes. • Transparency-Principle: Balancing usability with transparency to avoid overwhelming or misleading users about the complexities of AI mechanisms.	develop software applications. Focus on reducing manual coding effort, reusing existing code, and customizing applications for specific customer needs.	experimentation with novel AI technologies to sustain market leadership. Integrating innovation with customer-centric software solutions. Experimentation with novel AI models to maintain a competitive edge.	mechanisms and ensure employees critically reflect on AI outcomes. #2 Shared responsibility vs. clear allocation of individual responsibility: When building on open-source architectures, accountability may be distributed among open-source providers (e.g., OpenAI), the AI development company, and the end users, particularly in cases of bugs or system failures.	transparency about AI functionalities. • Establish Clear Contracts: Develop contractual agreements that clearly define accountability and responsibilities among all parties, including open-source providers and end users. • Political Initiatives: Engage in political initiatives to strengthen industrial legal standards and accountability practices in AI development and application.
L1 (AI Product Lead, Medical Technology Services)	• Strategic alignment: The organization has a clear focus on using AI and recognizes the importance of effective management even after AI implementation. • Knowledge: There is high AI-related knowledge among data scientists but they also need to have a certain level of domain knowledge to ensure effective collaboration and results. • Culture: The organization fosters an open culture characterized by active listening,	• Data: Own data management department for data analyses and forecast • IT Resources: Well-equipped organization (good data infrastructure and tools).	• Sustainability-Principle: Optimize manufacturing processes to minimize CO2 emissions.	Smart Manufacturing & Anomaly Detection: Utilize data analytics to optimize machine usage in production, identifying and addressing anomalies to enhance efficiency and reduce downtime.	The organization focuses on exploitation strategies, leveraging data analytics to optimize processes and improve operational efficiency.	Standardization vs. personalized approaches: Managers may prioritize protecting their personnel over fully optimizing processes through automation, fearing potential job losses and the impact on team morale, even if automation could enhance efficiency. This highlights the tension between technological advancement and safeguarding employee roles.	• Assign key user: Assign a key user responsible for overseeing the AI tool. • Ongoing management: Effective management remains crucial even after the AI system is implemented. • Quality validation: Implement quality management practices to validate the AI outputs. • Assign responsibilities: Share successes while clearly assigning responsibilities for each aspect of the process

	accountability, and a strong tolerance for failure. It embraces agile methodologies and encourages collaboration through cross-functional teams.						
E1 (Software Developer, Software Company)	• Strategic Alignment: AI should only be implemented by customers when it can effectively relieve personnel of repetitive tasks. Customers are ready to adopt AI when they recognize its potential to address their needs. • Knowledge: Organization has an open mindset and strong communication skills are essential, along with economic understanding to evaluate whether AI implementation is feasible and beneficial. • Culture: Organization as service provider has a shared open mindset. Employees must remain patient as AI models become reliable over time. Transparent communication throughout the process is key.	• Data: Service provider suggests that AI implementation is only meaningful when the use case occurs regularly, providing sufficient data to train and optimize models. Data quality is crucial, but data availability can be hindered by silos within organizations. • IT Resources: Software developer suggests that organizations need not only financial and personnel resources but also access to relevant data to support the AI use case.	• Data Security: Some organizations are hesitant to adopt cloud services due to concerns about potential data loss. • Transparency-Principle: Transparency issues may primarily affect customers, who might not fully understand how their data is being used or processed. • Fairness-Principle: AI tools could potentially discriminate based on historical data or inherent biases in human decision-making, leading to unfair outcomes.	Customer Service in Finance: The AI tool records and transcribes customer requests, such as inquiries about bank card blocking, and automatically initiates the relevant process, streamlining customer service operations.	The organization has a focus on exploitation strategies by providing AI-based customer support.	#1 Effective use of AI's capabilities vs. ethical violation: While AI has the potential to increase efficiency, customers may become highly dependent on AI providers due to a lack of in-house skills and resources, leading to increased costs from hiring freelancers. #2 Effective use of AI's capabilities vs. ethical violation: While some individuals will report loopholes to prevent misuse, others may exploit them for personal gain or to bypass systems.	• Regulations: Regulations should be tailored to the specific industry and use case, with sectors like finance requiring stricter oversight. • Pilot Testing: Conduct pilot testing to ensure that AI is suitable for the intended use case and meets performance standards. • Data Security: Offer local data centers to mitigate data security concerns and ensure compliance with privacy regulations. • Ethical Behavior: The organization can learn from individuals' ethical understanding to shape more responsible AI practices. • Responsibility: To address responsibility issues, the organization encourages rapid learning and frequent testing, adopting a "fail fast" approach.
E2 (Data Scientist, Car	• Strategic alignment: The company prioritizes AI research and foster innovation.	• Data: A dedicated data management department ensures high-quality data and	• Regulation: Differences between Chinese law and European	AI simulation and image generation: Leveraging AI to create virtual rooms	Focus on exploration as AI innovation is central to the	Effective use of AI's capabilities vs. ethical violation:	• AI Guidelines: Develop and implement a comprehensive internal ethics guide to outline

Manufacturer)	AI-process fit is a key focus area, aligning technology adoption with operational needs. • Knowledge: There is a heterogeneous distribution of AI-related knowledge across departments. While IT teams possess advanced expertise, non-technical areas like marketing show lower levels. • Culture: Organizational culture varies significantly; some departments demonstrate openness to innovation, while others maintain a more traditional, risk-averse stance.	robust architectural frameworks, enabling seamless AI integration. • IT Resources: Advanced IT infrastructure supports the company's AI initiatives, complemented by substantial investments in cybersecurity to safeguard digital assets.	law as a multinational company (especially as the European regulation is stricter, often hindering innovation but on the other hand establish ethical guidelines). • Non-maleficence-Principle: Safeguarding intellectual property is critical, ensuring that sensitive information such as construction designs is not exposed through AI tools.	that facilitate interactive car design, offering a creative platform for innovation.	company's core business value, with a focus on experimentation in cutting-edge applications like virtual design simulations. Exploration extends to improving non-core processes (e.g., marketing) to optimize workflows and increase productivity, illustrating some levels of exploitation.	On the one hand, there can be external cybersecurity threats: The company faces risks from potential cyberattacks targeting AI tools used for developing innovative designs, which could lead to intellectual property theft. On the other hand there is the internal risk of unsupervised and thoughtless AI usage, whereby confidential company information might be inadvertently unloaded to external platforms, such as ChatGPT.	acceptable AI usage and safeguard against misuse. • Information Security: Enhance awareness programs focusing on cybersecurity and the protection of sensitive business information, particularly targeting non-technical departments. • AI Knowledge: Promote knowledge-sharing initiatives between AI-savvy teams (e.g., IT) and those with limited AI expertise to bridge knowledge gaps and foster a cohesive approach to AI adoption.
E3 (IT Security, Manufacturing Company)	• Strategic alignment: There is limited strategic alignment with AI initiatives, as AI is not a central part of the company's vision. Efforts focus on reactive measures rather than proactive innovation. • Knowledge: The organization's AI knowledge base is minimal, relying more on individual employee efforts and private initiatives than	• Data: The company does not yet utilize its manufacturing data well for structured analysis or decision-making, indicating a lack of data-driven insights. • IT Resources Significant investments have been made in IT security infrastructure, but there is little to no funding allocated for AI development, pilots.	• Regulation: The global context and the autonomous actions of different AI bots that commission others make it non-transparent make it hard to do something against it • Non-maleficence-Principle: Ensuring that AI tools for cybersecurity do	AI bots used for cybersecurity and threat mitigation: The company recognizes that there are increasingly sophisticated, AI-driven hacking attacks. These AI tools focus on identifying blockers and loopholes to outsmart the blocking of malicious IP addresses. Thereby, the dynamic and self-learning nature of these attacks requires	The organizations follow exploitation: External pressures, such as AI-based hacking threats, drive the company to reluctantly adopt AI tools. Without expanding its knowledge and defenses, the company risks falling behind in	#1 Effective use of AI's capabilities vs. ethical violation: This dilemma surrounds the dual-use nature of AI tools, making hacking easier (by automating knowledge acquisition for malicious activities) that can also serve as critical defense mechanisms. AI used for defensive purposes learns how to block IP addresses, but adversarial AI evolves to bypass these defenses, creating a perpetual arms race. The company must	• Hands-on Training: Provide hands-on opportunities for employees to work with advanced AI-driven cybersecurity tools, encouraging both familiarity and expertise. • AI Guidelines: Establish internal guidelines that delineate the ethical boundaries of AI usage in defense mechanisms, ensuring transparency and compliance with industry standards. • Shift Focus: Shift a portion of IT security funding to proactive AI development and pilot projects to better anticipate and counteract autonomous hacking efforts.

	structured company training programs. • Culture: The organizational culture is restrictive and cautious, with a focus on using technology to safeguard resources rather than exploring AI's broader potential.		not inadvertently harm organizational assets or escalate security vulnerabilities.	constant evolution of the defensive	cybersecurity (punctuated equilibrium). The strategy is reactive, focused on using AI for immediate defense rather than proactively exploring AI's potential for broader applications.	continuously adapt to stay ahead. #2 Shared responsibility vs. clear allocation of individual responsibility: Who bears responsibility if defensive AI tools fail to secure the system, and what safeguards should be in place to prevent AI misuse?	
E4 (Recruiter, Headhunting Agency)	• Strategic alignment: The organization conducts experimental AI pilots to explore its potential but lacks a formalized strategy for AI integration. • Knowledge: Limited AI-related expertise within the organization, with no internal specialists or structured knowledge-sharing initiatives. • Culture: Open-minded organizational culture, emphasizing quality in candidate assessments and recruitment processes (with or without the help of technologies)	• Data: Candidate data is managed systematically but largely processed manually, limiting efficiency and scalability. • IT Resources: No dedicated IT infrastructure for AI adoption, which constrains the implementation of advanced tools.	• Non-maleficence-Principle: Issue of dehumanization (inappropriate or overly standardized assessments without empathy). Ethical concerns arise if AI is perceived as overly invasive or impersonal, particularly for high-level roles. • Fairness-Principle: Ensuring AI does not introduce biases in candidate evaluation.	Candidate assessment with AI: The organization has piloted the use of AI for personality testing and general interview questions in the initial stages of recruitment. The AI agent conducts preliminary assessments, with candidates progressing to in-person interviews for further evaluation.	The company employs AI experimentally, focusing on identifying areas where it can enhance decision-making without compromising quality (rather exploration).	#1 Effective use of AI's capabilities vs. ethical violation: Candidates may attempt to manipulate AI assessments by tailoring responses to what they believe the AI expects, undermining the integrity of evaluations. #2 Standardization vs. personalized approaches: The reliance on AI for initial assessments may conflict with the need for personalized engagement, especially for executive (C-level) positions, where candidates value human interaction. Similarly, recruiters may face role conflict, balancing AI-driven insights with their professional judgment.	• Human-in-the-Loop: For high-level positions (e.g., C-level), restrict AI to pre-screening candidates and leave detailed assessments to human recruiters. • AI Training: Provide recruiters with training on AI tools, including understanding AI outputs and addressing potential biases. • Communication: Transparently communicate to candidates about the role of AI in the assessment process, ensuring trust and understanding.
E5 (Software Developer,	• Strategic alignment: The organization is in the exploratory phase,	• Data: Manages a diverse range of customer databases;	• Regulation: GDPR compliance for	#1 Algorithmic platform for predictions:	Focus on exploration, experimenting	#1 Standardization vs. personalized approaches:	• Ethical Training: Educate employees and customers on the ethical implications of AI use,

Tax and Audit Consulting)	conducting AI feasibility studies to identify applications that align with business needs. • Knowledge: A highly skilled workforce with strong technical expertise and familiarity with state-of-the-art AI technologies. • Culture: A cautious approach to handling sensitive customer data, balanced with a commitment to adopting cutting-edge technology.	however, the data often lacks quality and consistency, presenting challenges for AI implementations. • IT Resources: Well-established IT infrastructure that supports both internal operations and customer solutions, including setting up technical environments for clients.	handling sensitive customer data. • Fairness-Principle: Ensure non-discrimination in data classification and automated suggestions. • Transparency-Principle: Maintain clarity in how AI processes and analyzes sensitive customer data.	Development of algorithmic platforms for data forecasting and classification. AI tools classify customers into operational categories and provide automated improvement suggestions for value optimization. #2 GitHub and Microsoft Copilot for programming and coding; Use of GitHub Copilot and Microsoft Copilot for efficient coding, programming, and automation of repetitive tasks, enhancing developer productivity.	with the latest AI technologies to maintain market competitiveness and drive innovation.	Employees demonstrate caution when handling company or customer data professionally but may fail to exercise similar prudence in their personal digital practices. This duality can lead to inconsistencies in data security attitudes and practices. #2 Trust vs. Overtrust: Customers may over-rely on AI-generated forecasts without critical reflection, unaware of potential limitations such as algorithmic biases or hallucinations.	emphasizing the importance of responsible data handling both professionally and personally. • AI Training: Train customers to recognize the potential for AI hallucinations or inaccuracies in forecasts, promoting a balanced trust in AI-generated outputs. • Legal Training: Conduct training to improve employees' and customers' understanding of data ethics and GDPR compliance.
E6 (Recruiter, Insurance Company)	• Strategic alignment: AI is primarily used unsupervised in areas like recruitment and content creation. The organization is exploring initial pilot projects to gain experience with AI technology (but so far, only conceptualizations). • Knowledge: Recruiters have moderate digital skills but limited AI expertise.	• Data: The organization is highly data-driven, but its data systems are not directly integrated with AI workflows. Rather data analysts as scientists. • IT Resources: A strong IT infrastructure exists, though it is not specifically optimized for potential AI deployment.	• Regulation: Adherence to legal requirements is prioritized. • Fairness - Principle: Ensuring AI tools treat all candidates equitably in recruitment processes. • Non-maleficence-Principle: Avoiding over-reliance on AI	#1 Candidate assessment with AI: Conceptual discussions on using AI tools for conducting personality tests and ranking candidates to streamline recruitment and save time and resources (currently the organizations solely uses an ABC-candidate classification). #2 Employer branding content	The organization focus on cost-efficiency, using AI to minimize operational expenses while maintaining competitive advantages (exploitation; supervised AI use). However, the unsupervised use of generative AI tools can be	#1 Trust vs. Overtrust: Without proper supervision, AI outputs may lack nuanced interpretations, whereby recruiters may overrely on AI, ignoring the impersonal nature of decisions and communications. #2 Standardization vs. personalized approaches: Challenges arise when AI-generated content becomes generic, reducing the company's	• Expectation Checks: Clearly communicate the role of AI in hiring to candidates and recruiters to build trust and manage expectations. • AI Guidelines: Develop and enforce guidelines for the unsupervised use of open-source AI tools, ensuring responsible and ethical application. • Value Alignment: Align generative AI tools with the company's broader strategy and tonality to ensure their effective use in creating unique value.

	• Culture: The company has a sedate, safety-oriented corporate culture, typical of the insurance industry (focus on risk mitigation).		tools that may erode creativity or personalization.	creation with generative AI: Leveraging ChatGPT and Dall-E for tasks like creating job offers and creating social media posts to enhance employer branding.	rather linked to exploration due to the medial hype surrounding AI (particularly ChatGPT; punctuated equilibrium).	ability to stand out, especially as competitors adopt similar tools.	
E7 (Recruiter, Technology Provider)	• Strategic alignment: AI is primarily used in recruitment processes and is mandatory for certain roles, focusing on optimizing hiring practices. • Knowledge: Recruiters have strong digital knowledge but limited understanding of AI's technical nuances and biases. • Culture: Internally moderately modern, with efforts to improve employee-facing systems. Mixed perspectives on the role of AI in recruitment.	• Data: Data management practices require improvement. • IT Resources: Focuses on innovation for customer-facing products but lags in adopting modern systems internally for processes like recruitment.	• Fairness-Principle: Aimed at reducing human bias but not completely bias-free. • Transparency-Principle: Candidates are informed about AI use and must consent. • Non-maleficence-Principle: Concerns about the dehumanization of processes, especially for empathy-driven roles like porters.	Pre-selection of candidates with AI: AI tools are used for pre-selecting candidates, particularly for trainee programs across different countries. AI handles tasks like personality tests and ranking candidates.	The organization adopts a dual approach to AI implementation, experimenting with AI tools internally—such as in recruitment—to enhance process efficiency and assess their reliability before deploying them externally for customer-facing applications. Internally, AI is leveraged to optimize and refine existing processes (exploitation), while externally, its focus shifts toward driving innovation and generating new business value (exploration).	#1 Organizational goals vs. personal employee needs: AI reduces personal interaction between recruiters and candidates, which can lead to a lack of empathy in assessments. #2 Trust vs. Overtrust: The performance in AI assessments may differ significantly from in-person evaluations, leading to mixed perceptions of AI's reliability among both recruiters and candidates.	• Non-AI Alternatives: Flexibility to adapt or omit AI assessments if candidates strongly oppose them. • Ethical Training: Ongoing training for recruiters to ensure critical reflection and ethical usage.

E8 (Data Analyst, AI Lab)	• Strategic Alignment: The managers of clients should reflect on the AI process fit and clearly define the intention behind AI integration. • Knowledge: The research department has strong technological skills, complemented by expertise in psychology and mathematics, enhancing interdisciplinary AI development. • Culture: The research department is highly technology-affine, fostering a culture of innovation and exploration.	• Resources: No own IT-resources as rather operating as consultant for businesses. Resources of the AI lab is personnel that possess high technological skills, ensuring the team is well-equipped for AI-related tasks. • Data: No specific data-related challenges identified in this scenario.	• Non-maleficence-Principle: There is a risk of becoming emotionally attached to a robot by forming a friendship, which may affect decision-making and ethical considerations. • Fairness-Principle: Hypothetically, an algorithm may discriminate against certain nationalities when evaluating criminality, so users must critically reflect on potential biases and ensure fairness in the system.	#1 Social Robots in care work: Robots designed for care work to provide social support, assisting individuals with daily tasks and offering companionship, especially in elderly care or rehabilitation settings. #2 Robots in manufacturing: Robots used in manufacturing environments that physically interact with humans and collaborate in tasks such as assembly, welding, or material handling, improving productivity and safety.	The research department focuses on exploration strategies by both teaching students and conducting research in the area of social robots, aiming to advance knowledge and innovation in this field.	Trust vs. Overtrust: Due to the physical appearance of robots in manufacturing, individuals may anthropomorphize them, overestimating their capabilities. Despite their human-like appearance, these robots are typically trained for specific tasks and cannot adapt their knowledge to new contexts or situations.	• Explainable AI: Explainable AI is critical for diagnostic detection systems to clarify where the suggestion or recommendation originates, ensuring transparency and trust in the AI's decision-making process. • User Engagement: It is important to get the user on board when developing social robots to avoid ethical concerns, ensuring their understanding and acceptance of the technology. • Learning by Doing: Adopting a learning-by-doing approach helps to better understand data and models, enabling users to actively engage with AI systems and address potential fairness issues.
E9 (Consultant, Business Analytics Institute)	• Strategic Alignment: AI readiness is seen as a prerequisite for learning, ensuring students are prepared to engage with AI concepts. • Knowledge: Students in the courses have no prior knowledge of AI, requiring foundational education in the field. • Culture: The students have limited prior	• Resources: A lack of financial resources and political support is perceived to hinder the development of AI literacy on a societal level. • Data: No specific data-related challenges identified in this scenario.	• Transparency-Principle: The provided student courses involve educating students on Explainable AI (XAI), fostering a deeper understanding of model transparency, which directly addresses transparency	Educate students on AI: The provided students courses involve educating students on AI by teaching Explainable AI (XAI) and AI governance mechanisms (including current regulations), allowing them to understand how training data influences AI outputs, fostering a deeper	The research department focuses on exploration strategies by teaching students at schools through innovative learning platforms that help them understand AI's functioning and its applications.	#1 Standardization vs. personalized approaches: While AI can enhance organizational efficiency, it may also lead to discrimination against individuals, potentially impacting fairness and equality in decision-making processes. #2 Effective use of AI's capabilities vs. ethical violation:	• AI Training: Implement AI training in school curricula and organizations to teach both its potentials and limitations. • Responsibility Awareness: Raise awareness that AI is merely a tool to avoid uncertainties in responsibility allocation. • Guidelines and Compliance: Political guidelines are essential, but organizational compliance is equally crucial. • Human Autonomy: AI use should not impede human autonomy.

	knowledge of technology, presenting a challenge for integrating AI into their learning experience.		issues by enabling students to critically analyze and interpret AI decision-making processes.	understanding of model transparency, and on how to develop models according to law.		If organizations have no own AI guidelines, there is a minimum of legal guidance required. However, too much regulation can hinder innovation and too much explainability mechanisms can hinder systems usability.	• System Design: Systems should be designed to prevent misuse and ensure ethical applications. • Societal Guidance: Society will guide moral values and help avoid ethical scandals through reputation loss.
E10 (Data Scientist, Software Company)	• Strategic alignment: Positioned as experimental leaders with a strong focus on innovation and customer-oriented solutions. Customer acceptance of AI is generally positive but tempered by concerns about AI's "black-box" nature and transparency issues. • Knowledge: Programming and technical expertise are core competencies within the organization. • Culture: A culture that embraces critical reflection, encourages learning from failure, and prioritizes solving complex problems through experimentation and adaptability.	• Data: Does not own proprietary datasets but possesses strong expertise in managing and architecting data systems. • IT Resources: Combines in-house tools with open-source platforms to maximize flexibility and innovation.	• Regulation: Lacks comprehensive legal frameworks to define accountability and liability in AI use. • Accountability-Principle: Responsibility for AI-related outcomes should primarily rest with those who develop and provide the tools.	Customer chatbot: Pilot implementation of a chatbot for 1st-level customer support, handling tasks such as communication, newsletter dissemination, order confirmations, etc. (built on GPT-based architectures).	The organization focus on exploration strategies for customer chatbots, experimentation is focused on delivering novel services and expanding the portfolio of AI-driven applications. For coding and programming, the emphasis is on exploiting AI tools to enhance productivity.	Shared responsibility vs. clear allocation of individual responsibility: The organization has deep knowledge of AI limitations (e.g., hallucinations or errors in outputs), but customers may lack this awareness, raising questions of who is responsible for potential errors or misuses of the technology (particularly, if the customer do not report when there are damages through the AI tools; not feedback processes to the software company).	• AI Training: Provide training in prompt engineering and effective interaction with AI systems to help customers better understand how to use the technology responsibly and interpret outputs. • Feedback Loops: Provide feedback mechanisms to customers to be informed when there are potential negative outcomes. • Transparent Design: Strive for a balance between transparency and usability, offering clear but accessible explanations of AI capabilities and limitations to customers.
E11 (AI Researcher, University)	• Strategic Alignment: The organization is customer-oriented, with	• Resources: No details on IT resources at the department, however, the personnel had a	• Transparency-Principle: When AI functions as a „black box,” it	Anomaly Detection & Intelligent Manufacturing: Use AI to monitor machines in real-time,	The organization focuses on exploration strategies by	#1 Trust vs. Overtrust: Different understandings of AI lead to varied usage: some focus solely on its advantages or	• Human Oversight: Humans must remain in the loop to prevent incorrect AI performance and ensure accurate decision-making.

	problems arising from the customer side, creating high AI business potential. • Knowledge: Diverse expertise is required, as no single individual can possess comprehensive knowledge across all areas of AI. However, personnel need a shared understanding of AI definitions and their respective roles within the organization. • Culture: The organization is open to new technologies and actively explores research on enhancing digital competencies.	shared understanding of AI and their respective role to incorporate their work with organizations. • Data: There are differing understandings between the organization and its providers regarding data quality and flow.	hinders the clear allocation of responsibility, making it difficult to trace decisions and hold individuals accountable.	detecting anomalies to prevent damage such as cracking, ensuring smooth operations and reducing downtime.	offering a new-to-market analytical tool for predictive maintenance, aiming to innovate and improve maintenance processes through advanced data-driven insights.	disadvantages, while others view it as an analytical tool and are more interested in understanding its functioning and underlying processes. organizations want to upskill vs. different role understandings #2 Standardization vs. personalized approaches: Relying on external providers for expertise creates a dependency, which may limit the organization's ability to develop in-house capabilities and increase long-term reliance on external support.	• AI Limits & Processes: It is important for organizations to define AI's limits and establish clear processes to manage its use effectively. • Legal Risk in Other Countries: In some regions, relying solely on AI may risk legal issues, which are only addressed after problems arise. • Reputation Impact: There is hope that the fear of reputation loss will encourage organizations to behave more fairly—e.g., if Uber discriminates, users may stop using the app. • Learning from AI: It is possible that individuals can learn from AI, improving their decision-making and skills based on AI-driven insights.
E12 (Innovation Architect, Venture Capital)	• Strategic alignment: The organization lacks clear guidelines for utilizing AI tools and interpreting their outputs effectively. • Knowledge: The organization demonstrates basic digital literacy and a willingness to experiment with new approaches. • Culture: There is a strong openness to adopting new technologies and a mindset focused on learning by doing.	• Data: Concerns exist regarding the quality and, at times, the availability of data. • IT Resources: While the organization has limited skilled personnel, there is a notable level of technological affinity and openness, resulting in a reliance on external IT providers.	• Fairness-Principle: Evaluate the improvement in performance by critically examining the underlying values embedded in AI systems. Reflect on whether these values align with organizational principles, ethical standards, and broader societal expectations.	#1 Information search with ChatGPT: Generative AI tools are used to estimate the size of the elderly care market. These tools provide a competitive advantage by improving efficiency and streamlining the analysis process. #2 Matching tool in real estate market: A matching tool is used to pair houses with customers. It streamlines the	The organization leans towards exploitation strategies, using AI to enhance existing processes and outcomes. For instance, AI is deployed to improve tasks such as information search, matching houses with customer preferences, and	#1 Trust vs. Overtrust: AI can provide access to areas where users have limited knowledge, such as market sizing for the elderly, but this can hinder critical reflection on the results, as users may overly trust the AI without fully understanding the underlying assumptions or potential biases. #2 Standardization vs. personalized approaches: While AI enhances efficiency and performance,	• Learning by Doing: Individuals will learn to use tools through trial and error, gaining familiarity over time. • AI User Guidance: AI should focus on guiding users through effective communication and asking relevant questions, prioritizing user support over full understanding of its underlying workings. • Pre-Implementation Testing: In cases where responsibilities between the user and provider are unclear, it is critical to test the AI beforehand to ensure it functions as intended and to build the provider's confidence in its reliability.

	However, the overall technological understanding among staff remains limited, with a preference for incremental innovation.			matching process and increases efficiency by automating the collection of customers' preferences and aligning them with suitable housing options.	content creation, emphasizing efficiency and incremental innovation over exploratory approaches.	organizations must reflect on whether the AI's actions and outcomes align with their moral values and ethical standards.	• Provider Responsibility: The provider remains responsible for failures resulting from logical or foreseeable issues within the AI system. • Clear Communication from Leadership: Leadership must clearly communicate how and where AI is being utilized within the organization to ensure alignment and transparency.
E13 (AI Researcher, University)	• Strategic Alignment: A study course is designed to help students become familiar with AI, although the current study case may not yet have significant business potential. • Knowledge: The department has a foundational understanding of AI, including programming skills, ethical considerations related to AI autonomy, and knowledge of data security. • Culture: The department is tech-savvy and eager to explore the latest advancements in AI and related technologies.	• Data: Department struggles because data security concerns can limit data availability, which is crucial for effectively training AI models. • Resources: As AI models grow in complexity, students sometimes lack the necessary resources and appropriate hardware. Political support is essential to ensure the institution remains competitive on a global scale.	• Data Security: While data security is not inherently an AI-specific issue, it should be considered from an ethical perspective to ensure the protection of sensitive information and the responsible use of data.	Mood Detection through face analysis: A university project where students develop a deep learning model to analyze facial expressions and detect the mood of individuals. The system uses computer vision techniques to interpret emotional states based on facial features.	The research department emphasizes exploration strategies, focusing on familiarizing students with AI technologies.	#1 Shared responsibility vs. clear allocation of individual responsibility: Germany risks falling behind the USA and China in AI development due to insufficient financial support from its government, while the USA benefits from intransparent models that are not fully understood or accessible to Germany, hindering its ability to compete effectively. Thus, dilemmas can arise when different moral perceptions on responsibilities and cultural norms regarding accountability between communities clash together #2 Shared responsibility vs. clear allocation of individual responsibility: In a hypothetical scenario of autonomous driving, it is difficult to define who	• Human Responsibility: Humans must retain accountability, supported by control mechanisms to determine whether responsibility lies with the system or the user. • Technical Safeguards: Implement robust technical solutions to minimize the risk of exploiting system loopholes.

						is responsible in case of an accident, but it should be ensured that humans can intervene if necessary to maintain control and accountability.	

Paper X

Stolz, K.; Hammerschmidt T. & Posegga, O. (2024). *How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers Toward Sustainable Consumption: Exploring Nudges for Green Decision-Making* (Best Paper Nominee). Proceedings of 57th Hawaii International Conference on System Sciences (HICSS), Oahu, Hawaii, <https://hdl.handle.net/10125/106970>.

How Conversational Agents Influence Purchase Decisions of Online Fashion Shoppers toward Sustainable Consumption: Exploring Nudges for Green Decision-Making

Accepted Version

Katharina Stolz
University of Stuttgart, Germany
katharina.stolz@bwi.uni-stuttgart.de

Teresa Hammerschmidt
University of Bamberg, Germany
teresa.heyder@uni-bamberg.de

Oliver Posegga
University of Bamberg, Germany
oliver.posegga@uni-bamberg.de

Abstract

With nudges, conversational agents (CAs) can be used to recommend environmentally sustainable products to individuals shopping online. CAs can thus influence individual purchase behaviors and have the potential to promote green decision-making. There is a lack of qualitative insights into how CA nudges might influence the purchase decisions of individuals in the specific context of sustainable fashion consumption – especially regarding customer perceptions of CAs trying to influence those decisions. We conducted an explorative survey with a qualitative online questionnaire of 79 fashion shoppers to determine how they think about CAs nudging their product choices and to derive propositions on how CA nudges should be designed to support green decision-making.

Keywords: digital nudges, recommender systems, fashion industry, sustainable consumption, green IS

1. Introduction

Fostering environmental sustainability and using digital technologies to promote sustainable shopping behavior of customers are among the fashion industry's top challenges today (Kim & Kim, 2020) and are part of the ongoing debate on “what consumers and suppliers can do” (Stöckigt et al., 2018, p. 188) to enhance sustainable consumption (Pero et al., 2020). Sustainable consumption aims to balance current customer needs with reducing environmental pollution (WCED, 1987). One way fashion providers can foster environmental sustainability is to use digital technologies that guide customers toward more environmentally sustainable product choices during their online shopping (Pero et al., 2020).

Conversational agents (CAs) such as chatbots are particularly appropriate for this task; fashion shoppers can use them to make shopping-related requests, optimize product searches, and provide feedback (Zhang & Balog, 2020, p. 1512). CAs interact with users by using “written or spoken natural language” (Diederich et al., 2019, p. 1550) to achieve a natural style of communication (Montenegro et al., 2019; Rheu et al., 2021). They represent a class of recommender systems that can influence user behavior (Cacanindin, 2020; Sutanto et al., 2021) by offering information that guides users to make environmentally friendly decisions (Sutanto et al., 2021).

Convincing customers to make environmentally sustainable product choices is challenging, as sustainable products are often more expensive and involve certain restrictions (Pero et al., 2020; Sutanto et al., 2021; Xiao et al., 2022). Hence, nudges delivered through a CAs' interface to guide user behavior in a specific direction are helpful to influence shopping behavior (Caraban et al., 2019). A nudge can be defined as “any aspect of the choice architecture that alters people's behavior predictably without forbidding any options or significantly changing their economic incentives” (Thaler & Sunstein, 2009, p. 6). Nudges based on Artificial Intelligence (AI) offer great potential to make customers think about their shopping behavior by heightening customer consciousness of sustainable fashion through human-like conversation (Yadav & Pathak, 2017); they can also use past customer transaction data to predict individual preferences and make individualized purchase recommendations (Kumar et al., 2019). Hence, nudges can make CAs more persuasive, and AI can improve the persuasive power of CA nudges.

An initial screening of CAs of the 50 most valuable fashion providers (Brand Finance, 2022), as well as the ten online fashion platform providers with the highest net sales in Germany (Statista, 2021) between October 2022 and June 2023 shows that there is great potential for CA improvement to foster sustainable consumption in the real world. Of these fashion providers, only five offered CAs that are capable of conversation beyond providing standard replies to simple questions (Calvin Klein, Michael Kors, Dior, H&M, and UNIQLO). In April 2023, Zalando announced it would soon launch an AI-powered fashion assistant based on ChatGPT (Zalando, 2023). When simulating typical conversation scenarios between fictitious customers and those five CAs (e.g., customers asking, “How is this product produced?” or “Are there certain quality labels or hallmarks?”), the CAs' answers did not include aspects of sustainable consumption. Asked whether there are fair-traded or fair-produced products, none of the CAs provided a suitable response. Answers to questions about whether the fashion providers follow certain guidelines or engage in environment-friendly production practices were more on target (e.g., “At UNIQLO,

we strive to change the world for the better through the power of clothing. For more information on sustainability and ongoing projects, please visit: [URL];” “Sustainability and the protection of our environment is more than a matter of course. For my employer, it is a matter of the heart. –Otto”).

While there are studies of how CA nudges can influence customer green decision-making in the food (Sutanto et al., 2021) and tourism (Um et al., 2020) industry contexts, such research within the fashion industry is rare. The observations above are only preliminary; real cases for more in-depth study are lacking. Empirical research is required to understand better how CA nudges can be used to drive green decision-making by fashion shoppers. In addition, researchers have called for more qualitative investigations of green consumer behavior, from the customers’ perspective, within digital contexts (e.g., Sutanto et al., 2021; Xiao et al., 2022). Therefore, we explore online fashion shoppers’ perceptions of CAs trying to influence their shopping behaviors by nudging them toward purchasing more environmentally sustainable fashion products. Further, we derive propositions on how CA nudges should be designed to support green decision-making by online fashion shoppers. Our research question (RQ) is: *How do online fashion shoppers perceive CAs trying to influence their purchase decisions toward green decision-making?*

Given that our RQ aims to explore the potential of CAs to convince customers to choose sustainable fashion products and, thus, concerns a highly subjective and individual topic, we opt for an explorative approach (Sarker et al., 2018). Based on a qualitative online questionnaire with 79 fashion shoppers, we explore their perceptions regarding CAs trying to influence their product choices when shopping online. From that, we derive propositions on how to design CA nudges that foster sustainable fashion consumption.

The next section presents the theoretical background and our methodology. We then present our results in the discussion, followed by a conclusion.

2. Theoretical background

In this section, we develop the theoretical background used to guide the design and execution of our exploratory study, following Sarker et al. (2018).

2.1 Recommender systems: fashion CAs

In the fashion industry, recommender systems are advantageous because customers can choose from millions of pieces online (Deldjoo et al., 2023). Effective recommender systems are crucial for successful e-commerce that supports customer shopping procedures (Chakraborty et al., 2021).

Recommender systems, which are ubiquitous in e-commerce (Sysko-Romańczuk et al., 2022), can be described as “software tools and techniques providing suggestions for items to be of use to a user” (Ricci et al., 2010, p. 1). Thus, they assist in making decisions among various alternatives (Sysko-Romańczuk et al., 2022) by sorting information and making it available to users, which increases shopping transparency and efficiency (Chakraborty et al., 2021).

CAs that aid customers in supporting search, providing feedback, and disclosing preferences are specific types of recommender systems (González-García et al., 2023; Rima et al., 2023; Zhang & Balog, 2020). Recommender systems in the fashion industry are characterized by the fact that they operate in a very dynamic environment and require a highly specific vocabulary (Deldjoo et al., 2023). Also, due to seasonal product changes, CAs need to be very adaptable (Hwangbo et al., 2018).

Recommender systems, particularly those that use AI, can provide customers with personalized recommendations based on their individual preferences and characteristics. Thus, they offer exclusivity, which is especially crucial in the highly competitive fashion market (Moon et al., 2017). Recommender systems are thus an innovative solution to overcome the limitations of e-commerce, particularly the absence of physical staff and personal face-to-face advice (Hwangbo et al., 2018).

To avoid bias in survey participants’ answers, our qualitative online questionnaire uses the term CA and does not further differentiate between CA types (e.g., voice- or text-based CAs) or the underlying technology of the CA (e.g., AI-based or not).

2.2 Green IS, persuasive IS, and decision-making

Our work focuses on making specific electronic markets (e.g., fashion markets) more sustainable by reflecting sustainability within the interaction between customers and providers. This can be aligned with research on green IS concerning, *inter alia*, making technology-mediated processes more sustainable by investigating how technologies can impact market mechanisms (Alt, 2020). Following Alt (2020), there are several perspectives on green IS. On the individual level, it is about using IS to promote individuals’ green decision-making (Corbett, 2010, 2013).

Decision-making is the act of building a preference out of several alternatives (Orlovsky, 1978), which involves making a choice (Sadovykh & Sundaram, 2017). In the model of Mintzberg et al. (1976), the decision-maker must first recognize a problem or a tangible issue that requires action to begin evaluating alternatives (Sadovykh & Sundaram, 2017). Hence, using CA nudges to make fashion shoppers more aware of sustainability issues can be required, especially among individuals unaware of the environmental pollution debate (Sutanto et al., 2021). Green IS research, therefore, involves consideration of how to foster thinking about sustainability thoughts in the choices of individuals in order to achieve green decision-making (Corbett, 2010).

Given that green IS aims to motivate sustainable behavior, it is persuasive by nature (Corbett, 2010, 2013; Oinas-Kukkonen & Harjumaa, 2009). A persuasive system is “designed to reinforce, change or shape attitudes or behaviors or both without using coercion or deception,” indicating that persuasion aims to influence individuals’ thoughts, feelings, and/or actions (Oinas-Kukkonen & Harjumaa, 2009, p. 486). IS, primarily digital and mobile IS, such as CAs, offer the opportunity to provide individuals with information continuously, thus fostering persuasive interaction (Shevchuk & Oinas-Kukkonen, 2019). The persuasion context becomes relevant to change behavior and attitudes through persuasive IS (Oinas-Kukkonen & Harjumaa, 2009). Following Oinas-Kukkonen and Harjumaa (2009), the persuasion context consists of the following:

- the intent (involving the persuader and the type of behavior change),
- the event (involving the use, the user, and the technology), and
- the strategy (involving the message and route).

In our work, the persuader is characterized by the fashion providers giving access to the CA via their web shops or shopping platforms. Sustainable product choices, such as customers’ green decision-making, characterize the type of behavior change. In contrast, online shopping for fashion products is the underlying use context. The CA and its features represent the technology (which can be aligned to green IS and persuasive IS). The message is the concrete action through which changes in the behavior should happen, such as the dialogue of the CA interacting with the customer. The direct and indirect processes present the route to communicate the message, which can be represented through several categories of nudges that should influence people interacting with the technology (Caraban et al., 2019).

2.3 CA nudges

According to the aforementioned definition in the introduction, a nudge comprises the influence of the choice architecture without prohibitions regarding certain options and no significant changed monetary incentives (Thaler & Sunstein, 2009, p. 6). Nudges are particularly effective when individuals are uncertain and unclear regarding their preferences (Caraban et al., 2019).

Hansen and Jespersen (2013) divided nudges into four categories based on the two cognitive modes of thinking that lead to an action of the user, which the authors derived from dual process theory from Thaler and Sunstein (2009): Type 1: automatic thinking (e.g., uncontrolled, effortless, associative, unconscious) vs. Type 2: reflective thinking (e.g., controlled, effortful, deductive, slow, self-aware, rule-following); and the two levels of transparency of a nudge (transparent nudge vs. non-transparent nudge).

Based on these categories, Caraban et al. (2019) identified six categories of CA nudges, which can be structured using the dimensions “transparency and reflective” vs. “automatic,” as originally discussed by Hansen and Jespersen (2013). In the following, we define the nudges in the context of sustainable fashion:

- *Facilitate*: diminishing one’s mental and physical efforts. For instance, nudging customers by presenting sustainable products as a default option and making it more difficult to search for fashion products that contribute significantly to environmental pollution.
- *Confront*: the effort to stop an unwanted action through nudges that create friction, such as asking customers whether they are sure about buying products from fashion brands that do not have green labels (such as fair-trade labels, child-labor-free labels, and/or organic labels), and reminding customers of the consequences, such as calculating the CO2 footprint for the shipping of articles. To this end, several labels and/or government and NGO certifications, such as Fairmined, Global Standard gGmbH (GOTS), OEKO-TEX Service GmbH, and Swiss Better Gold Association, can be used in the fashion industry to increase transparency and validation that specific standards have been met (Haug & Busch, 2016; Osburg et al., 2021).
- *Deceive*: tricking users’ empathy and influencing how “alternatives are perceived” (Caraban et al., 2019, p. 6). For example, by making environmentally friendly alternatives more attractive by visualizing the greater number of trees that can be planted through the purchase of certain fashion products compared to other products.
- *Social influence*: using the aspiration of people to conform, such as raising the visibility of users’ actions and enabling comparisons (e.g., providing a waste balance of an individual user’s purchases compared to others).
- *Fear*: evoking negative feelings of loss or uncertainty, such as by making resources scarce to reduce the perceived availability of product alternatives.
- *Reinforce*: making individual behavior more prominent or provoking empathy (e.g., presenting pictures of environmental pollution) to encourage sustainable behaviors.

Facilitate nudges are primarily part of the non-transparent and automated dimensions spelled out above. The *confront* category can be classified as reflective and transparent, while those in the *deceive* category belong to the non-transparent and reflective dimensions. *Social* nudges are reflective and transparent, while *reinforcing* nudges are transparent and can either support reflective or automatic thinking. The category of *fear* is integrated mainly into transparent nudges and automatic contexts. It should be noted, however, that not all nudges belong to one dimension, nor can they permanently be assigned solely to one field of the matrix of Hansen and Jespersen (2013). Figure 1 is an overview. In the discussion section below, we use fashion CA nudges to provide a deductively derived structure for making propositions on the design of CAs.

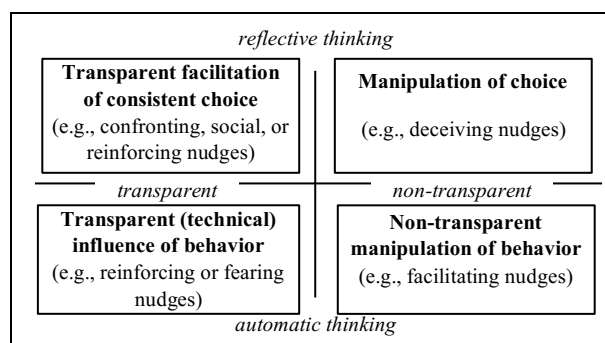


Figure 1. Categorization of fashion CA nudges.

Source: adapted from Hansen and Jespersen (2013) and Caraban et al. (2019).

3. Method

To answer our RQ, we conducted an exploratory study with a qualitative online questionnaire using open-ended questions. Exploratory studies can be mainly used for inductive reasoning (Sarker et al., 2018). We use the theoretical background above to guide the examination of our qualitative online questionnaire and provide a deductive structure for deriving propositions (see the discussion below). The exploratory focus allows us to investigate the impact of CAs from the customer perspective by exploring customers' lived experiences and expectations (Creswell & Poth, 2016). In the following, we describe the sample, questionnaire design, and analysis.

3.1 Sample

The survey was distributed between September and October 2022 through the platform Clickworker (77.22%) and via social media (22.78%). The channels were chosen because they allow for targeting participants with a preference for online fashion shopping and technology (Peer et al., 2017).

In the beginning, we provided all participants with initial information on our research project and conducted two questions for an aptitude check. First, we asked participants about their fashion shopping behavior to identify whether they qualify as online fashion shoppers. Second, we asked participants to describe their CA experiences in general to identify whether they had a basic affinity with CAs so they would be able to respond to our questions. Of 110 initial participants, 79 were usable and we excluded 31. In most cases, using targeting mechanisms to identify fashion shoppers worked well, and we excluded only 6 participants through our first aptitude check. The remaining 25 participants were removed because they reported no or only low levels of experience with CAs in the second aptitude check (n=17) or due to inappropriate answers (e.g., "...", n=8).

This results in a convenience sample with 79 participants (60.66% female, 37.70% male, and 1.64% not specified). Age ranged from 18 to 67 years, with a mean of $M=27.94$, $SD=10.00$. Some 13.42% of the participants had no higher education entrance qualification; 42.95% had a higher education entrance qualification as their highest educational qualification; and of all participants, 43.62% had attained a university degree. Some 54.10% answered the questionnaire in German and 45.90% in English. Of the 79, 70.89% are employees, 12.82% are students or trainees, 8.10% are freelancers, and 8.10% are retirees. Of those with jobs (n=56), 16.67% each work in consultancy, consumer services, and media, and another 14.58% each in manufacturing, consumer goods, or education. Some 6.25% work in other industries (e.g., healthcare).

Some 49.37% stated that they use CA frequently during online shopping (*regular users*, n=39), and 50.63% stated that they have some experiences with CAs during online shopping (*occasional users*, n=40). Among the regular users, we also identified 8 participants who reported working with chatbots for professional purposes, such as by providing technical support or designing chatbots (*specialized users*, n=8). In our results, we distinguish between these groups of *regular*, *occasional*, and *specialized users*.

3.2 Questionnaire

Since we aim to explore the perceptions of individuals regarding CAs trying to impact their purchase decisions during online fashion shopping, we captured the qualitative data of participants' experiences and opinions in the form of text responses to open-ended questions. This allowed participants to express themselves freely without restrictions, using their own words (Byrd et al., 1992). The absence of a human interviewer in online surveys who can clarify questions (Choy, 2014) led us to check the understandability of our survey through a pilot study with two individuals we acquired randomly via social media. The subjects of this pre-test were able to provide suggestions for changes to our questions, and we made subsequent changes to their wording.

The questionnaire begins with the two aptitude-check questions (see 4.1), followed by a request for sociodemographic data. The main part of our survey includes ten open-ended questions (see Table 1). Participants were informed that the questionnaire was set in a fashion context, eliminating the need to include the term "fashion" in every question.

Table 1. Qualitative online questionnaire.

Understanding and usage of CA in general	1. In which situations do you use CA in general (besides fashion shopping)? <i>Please describe situations when you use CA.</i> 2. What types of CA do you use in general, and for what purpose? <i>Please describe what CAs you use and why.</i> 3. How would you describe CA?
Experiences with fashion CAs	4. For what purposes do you use CA during online fashion shopping? <i>Please describe the purposes of CA usage during online fashion shopping.</i> 5. Have you already received a CA's suggestion to buy another fashion product during online shopping, and how did you react? <i>Please describe the situation and your reaction (previous experiences).</i> 6. When you buy fashion products online, what mechanisms of CA or information provided by the CA do you expect? <i>Please describe CA mechanisms/features/information that you expect (expectations).</i>
Perceptions and expectations of fashion CA	7. What do you think would affect your online shopping behavior? 8. What do you think about CA trying to influence your online fashion shopping behavior, and how do your perceptions change when the CAs' influence is intended to increase environmentally sustainable consumption? 9. How do your perceptions change when the CA is based on AI technology? 10. What are your expectations of CAs trying to influence your product choices for green decision-making during online shopping?

3.3 Data analysis

Given the exploratory nature of our survey, participants' answers regarding how they perceive CAs trying to influence their online shopping behavior for sustainability reasons were inductively coded based on (Mayring, 2014, pp. 80–87) and by using Microsoft Excel. Table 2 is a snapshot of inductive coding for the expectations toward CAs, whereby the quotes in the left-hand column are examples of the sorts of statements from participants that we placed in the right-hand column's category. The middle column is the core statement common to all statements assigned to a given category by the coders.

Table 2. Snapshot of inductive coding.

Quotes	Code	Category
"CAs should not influence shopping behavior because it would limit the perspective of customers to specific product classes." [I68]	CAs should not influence behavior	The ethical ideology of fashion shoppers
"Well I think personalized recommendations are a handy tool in the online retailer's toolkit. ... To me, I see ethics to be the responsible provision of service in line with laid down protocols." [I71]	CAs should be a standardized tool following standardized ethical protocols	
"Nudging behavior is only adequate if the nudging processes are communicated transparently." [I4]	CAs should solely use transparent nudges	CA ethics
"CAs should be ethical by design, meaning that they should be designed to be sensitive to values like security, safety and accountability and transparency." [I86]	CAs should follow certain ethical principles	
"CAs are essentially smart robots that are programmed to answer questions. They understand what you want and then give you the answer you are looking for. Intelligent conversational chatbots are built on machine learning and become more 'knowledgeable' the more you feed it data." [I90]	CAs can communicate intelligently	AI-based CAs
"A chatbot is a software or computer program that simulates human conversation or "chatter" through text or voice interactions." [I78]	CAs stimulate human-like behavior	

To categorize the answers of the participants into propositions on how fashion providers should design CA nudges to strengthen green decision-making, we used the nudging categories as introduced in section 2 above,

following a deductive coding scheme suggested by Mayring (2014, pp. 95–98). We coded in parallel and conducted several discussions among the coders until a consensus was reached to enhance the quality of coding. The statements were translated several times in both directions – English and German.

4. Results

- In this section, we present four main findings of our exploratory survey. We provide insights into the perceptions of our participants regarding CAs trying to influence their online shopping behavior.

4.1 Differences between user groups

- We identified that participants mainly interacted with text-based CAs, such as chatbots (57.69%), representing the persuasion element. The primary purposes of CA usage (persuasion intent) are direct (product) requests (85.57%), providing product feedback (8.65%), or using certain automated customer services (5.78%). Regarding other sociodemographic information, such as gender and age, we did not recognize any patterns.

- Comparing regular (n=31), occasional (n=40), and specialized (n=8) users, we identified that most of the **regular users** stated that they care about sustainability (64.00%). Users in this group stated that CAs that nudge customers toward sustainable shopping behavior would reinforce persons who already care about environmentally friendly shopping (19.67%) as it would help them find sustainable products quicker (8.41%) and would support their shopping procedures by providing more information and a greater variety of product alternatives (28.06%). As one said, “At least it would save the stress of searching for alternatives myself” [I68].

- None of the participants in the **specialized user** group supported the idea of using CAs to promote sustainable consumption, as this, they argued, is not the purpose for which they are designed (33.33%); they should focus instead on increasing shopping efficiency (66.67%) or providing product information (25.00%). As one said, “CAs should focus pragmatically on what they were designed for” [I8].

- None of the specialized users reported negative experiences with CA nudges. Most (75.00%) perceive CAs as modern and reliable information sources (25.00%) that can help fashion providers increase their customer base with sustainable shoppers. Therefore, whether participants in this group would follow CA nudges depends mainly on the appropriateness of the suggestions and how CAs can support the customer by saving time (enhancing the user experience). In other words, if “[CA] can give good advice” [I79] or provide “cost- and time-saving opportunities” [I90].

- **Occasional users** showed more dispersed results. Some stated that CAs trying to influence their shopping behavior is disturbing (23.81%), intrusive (9.52%), or a marketing trick to sell more products (9.25%). For instance, participants with CA experiences in which they did not follow the CA recommendation, stated that the recommendation appeared at an “inappropriate moment” [I1] when the reason for contact was unclear. Hence, participants perceived the suggestion as “a scam to sell products [that are] more expensive” [I11], “annoying” [I9], “manipulating” [I20], or “spam” [I57]. In comparison, others stated that CAs would optimize their shopping experiences (23.81%) and increase product information transparency (14.29%). For instance, one participant said, “I think this action is brilliant because I needed a replacement item that has similar item specifications to replace the item I ordered” [I53]. Hence, they see benefits of CAs when they offer great usability (52.05%) and follow certain ethical principles (50.98%), such as being transparent about their nudges (46.03%) and protecting customer data (50.70%).

4.2 CAs require intelligent abilities to support sustainable decision-making

- Some 16.46% of participants who affirmed that the ability of CAs to support sustainable decision-making mentioned similarities to human-like behavior (occasional users, 61.54%; regular users, 38.46%; no specialized users). For instance, one occasional user described CAs as “conversational software agents or chatbots, which are systems designed to communicate with human users by means of natural language often based on artificial intelligence” [I91]. Participants in other groups characterized CAs as “being respectful” [I32], “responding politely” [I83], “being friendly” [I49], and “allowing humans to interact with digital devices as if they were communicating with a real person” [I44]. Regarding feedback requests of the CA, one participant designing chatbots (specialized users) stated that “anthropomorphism as well as the need to stay consistent significantly increase the likelihood that users comply with a chatbot’s request for service feedback” [I90]. Participants often linked CAs with AI and especially the human-like (or anthropomorphic behavior) that AI can offer, given that AI would affect the ability of a CA to increase awareness of sustainability issues during online shopping (e.g., I18, I37, I44). However, most participants thought that responsibility for shopping decisions should remain with humans (84.81%). Specialized users who work with CAs in their professional capacities, in particular, did not agree at the same level (only 62.50% expressed that decision-making should remain the sovereignty of human beings). Among the reasons given were that “CAs can make intelligent analysis” [I84] and, therefore, they can

“actually make the better choices most times” [I89] than humans. Hence, AI might improve the ability of CAs to support sustainable decision-making of fashion shoppers, but it should not take over the final product choice.

4.3 Sustainability-related recommendations entail expectations of ethical compliance

- Participants expressed their openness to CA recommendations if the underlying nudging mechanisms are characterized by transparent processes (79.75%). As one put it, “Is it clear where the boundaries of the technology are? Is it clear what happens to the data?” [I18]. Another said, “It should be made transparent” [I62]. Participants related transparency to the declaration of the CA’s limitations (23.45%), such as “the limits in the program code” [I10]. Participants expressed high expectations for CAs to adhere to ethical principles (64.56%). For instance, participants argued that CAs should not discriminate and should have “a sound knowledge of the ethical structure of the respective culture” [I3] and should be “participative, [and] inclusive” [I16]. In addition, the information CAs present should be consistent: “It should be true to what it is standing for” [I48] and, to make the CA more reliable, humans should be available in the background to intervene and provide personal contact if the CA does not react adequately [I13]. Results further highlight the importance of providing background information on human rights in the manufacture of fashion products (11.39%) in addition to information regarding environmental pollution (7.59%), for instance, by illustrating the (non-)existence of green labels. Participants from the specialized user group, though, argued that the ethics of CAs are limited to the designer’s own sustainability behavior (20.00%). This might be one reason why this group expressed the lowest requirement that CA design principles adhere to certain (moral) standards. In general, the high expectations of CA ethical compliance among our participants (regular and occasional users in particular) may be due to the topic of interest (sustainability). Thus, using CAs to support green decision-making also involves considering ethical concerns.

4.4 Moral ideologies moderate CA nudges

It should be noted that most of the participants supporting the use of CA nudges to foster sustainable consumption generally perceive sustainability as essential for society (especially those in the regular user group; see section 4.1) – even though not all reported sustainable behavior during their own online shopping. For some, CAs offer the possibility to raise their awareness of sustainability issues when they are shopping online (21.61%). Our results highlight that whether CAs can drive sustainable consumption is affected by the moral ideology of customers. For instance, one participant stated, “I am very environmentally conscious and value sustainable consumption. When I shop, I look for organic products and eco-labels, and I inform myself about brands” [I12]. Another stated, “I always prefer products that are more sustainable” [I39]. Hence, the moral ideology of customers, in particular, affects how specific nudges might be appropriate to increase the likelihood of sustainable consumption via CA interaction.

5. Discussion

Based on our empirical results, we use insights into how our participants perceive CA trying to influence purchase decisions to derive propositions for how online fashion providers can design CA nudges to support green decision-making. We use the results in section 4 above to structure and derive these propositions, in combination with our theoretical background, within three areas for future research.

5.1 Ethical CA design

Irrespective of participants’ perceptions of using nudges to foster sustainable consumption, participants called for adherence to certain ethical design principles. Hence, an ethical CA design might strengthen the willingness of the participants to cooperate with the CA. Without ethical CA design, participants are not willing to follow CA nudges. Such an ethical CA design must be considered in all six categories of nudges and support a “transparent influence of behavior” following Figure 1, but not directly impacting sustainable consumption. Therefore, our results highlight ethical considerations with respect to the technology’s risks in addition to looking at the potential of using CAs to support green decision-making of online fashion shoppers. CA designers must comply with the ethical standards expected by online shoppers. Future work should examine the role of ethical CA design as an antecedent that enables CA interaction as a starting point for supporting sustainable consumption, as we assume a moderating influence.

Proposition 1: *Ethical CA design is an antecedent for using CAs to guide product choices toward sustainable decision-making of online fashion shoppers.*

5.2 Specific CA nudges address specific sustainability purposes

Results highlighted that CA nudges can reinforce people who already demonstrate sustainable shopping behavior as well as increase the awareness of those who consider sustainability to be a relevant topic but do not yet reflect that awareness in their own shopping behavior to a significant degree. This supports previous research arguing that sustainable consumption is affected by the moral values of individuals (e.g., Sutanto et al., 2021). Some participants expressed that CAs recommending environmentally friendly alternatives would support their knowledge of sustainable products and, thus, reinforce their shopping behavior mainly through nudges that strengthen automated thinking (e.g., reinforcing nudges). Those participants stated that they are already sustainable shoppers who pay attention to the sustainable use of resources. Other participants, though, stated that CA recommendations would make them reflect on their shopping behavior as they care less about sustainability while shopping. To make online shoppers question their behavior requires nudges that foster reflective thinking (e.g., confronting nudges). In general, participants demanded high ethical standards of CA nudges, which speak against using intransparent and manipulation nudges (e.g., fearing nudges). Hence, CA designers must consider that nudges supporting reflective or automated thinking should be communicated transparently. Further research is needed to understand differences in the moral ideology of individuals and how this affects the appropriateness of different CA nudges. In addition, researchers should ask participants about their moral ideology when conducting studies concerning sustainable consumption as a way to gain more in-depth insights.

Proposition 2: *The effectiveness of CA nudges to guide product choices toward sustainable decision-making depends on the moral ideology of fashion shoppers.*

5.3 AI increases the nudging potential of CA recommendations

As in earlier literature (e.g., Ghandeharioun et al., 2019), our results highlight that AI could make a natural style of communication more likely, foster human-like conversation with CAs, and thus help to establish trust through design (Rheu et al., 2021). In doing so, AI can help spur behavior changes in individuals interacting with CAs for fashion shopping (Deldjoo et al., 2023). Participants saw AI as relevant in influencing the ethical behavior of customers appropriately during human-CA interaction by making individualized, nuanced, and specified suggestions through AI's ability to learn during the conversation based on the customer's data. They assumed that AI is the underlying technology of the CA that enables it to make appropriate and customer-friendly suggestions and, thus, provide noticeable advantages during online shopping. Interestingly, participants also stated that responsibility for final shopping decisions should remain with humans. Hence, CA designers should be aware that AI has a mediating potential to strengthen the impact of CA nudges. Future work should investigate how online fashion shoppers trust or distrust AI-based CA recommendations compared to non-AI-based CA recommendations to identify the role of trust in AI for CA recommendations. Even though recommender systems are typical for fashion (Chakraborty et al., 2021), we think that trust in AI might be crucial – irrespective of whether the persuasion context is fashion shopping or other online shopping scenarios.

Proposition 3: *Trust in AI leverages the nudging potential of CA recommendations aimed at fostering sustainable consumption.*

6. Conclusion

Using CAs to guide sustainable decision-making has tremendous potential. With empirical data from 79 participants of a qualitative online survey, we gained initial insights into how participants perceive CAs trying to influence their purchase decisions regarding environmentally sustainable fashion products.

The explorative qualitative research design comes with some limitations. *First*, it would be important to extend the exploratory work with experimental designs to detail our insights, using concrete CA examples and actual CA-customer interaction. This is because the qualitative online survey is based on the self-appraisal of participants without concrete scenarios in which customers interact with CA to test and validate the nudges' impacts empirically. Future researchers could, for example, conduct an experimental study and investigate actual CA-customer interaction. Differently worded nudges (facilitate, confront, deceive, social influence, fear, or reinforce) could be used to deepen the understanding of green nudges.

Second, the propositions presented are only an initial insight into the acceptance and effectiveness of nudges in the context of CAs; they should not be considered conclusive. Extending the propositions is of great importance for the theory as well as for the design and use of CAs in practice. Future studies could, therefore, investigate additional influencing and contextual factors, which could be presented in a detailed framework or model.

Third, in alignment with similar work in the food consumption context (e.g., Sutanto et al., 2021), future work could elaborate on how the moral ideology of customers affects sustainable product choices and how this differs between several consumption contexts (e.g., beauty, fashion, food). In addition, a multi-stage sampling process could be applied to include a more diverse and representative sample, which would enhance the reliability and validity of the findings.

Finally, we hope this work provides valuable guidance for future research on fostering sustainable fashion consumption to reduce environmental impact.

7. References

- Alt, R. (2020). Electronic markets on sustainability. *Electron Markets*, 30(4), 667–674.
- Brand Finance. (2022). *Luxury & premium 50 2021 ranking*. <https://brandirectory.com/rankings/luxury-and-premium/table>
- Byrd, T. A., Cossick, K. L., & Zmud, R. W. (1992). A synthesis of research on requirements analysis and knowledge acquisition techniques. *MIS Quarterly*, 16(1), 117–138.
- Cacanindin, N. M. (2020). Greening food consumption using chatbots as behavioral change agent. *Journal of Advanced Research in Dynamical and Control Systems*, 12(01), 204–211.
- Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 ways to Nudgen. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Chakraborty, S., Hoque, M. S., Rahman Jeem, N., Biswas, M. C., Bardhan, D., & Lobaton, E. (2021). Fashion recommendation systems, models and methods: A review. *Informatics*, 8(3), 49:2-34.
- Choy, L. T. (2014). The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches. *IOSR Journal of Humanities and Social Science*, 19(4), 99–104.
- Corbett, J. (2010). Unearthing the value of green IT. In *Proceedings of the International Conference on Information Systems (ICIS)*.
- Corbett, J. (2013). Designing and using carbon management systems to promote ecologically responsible behaviors. *Journal of the Association for Information Systems*, 14(7), 339–378.
- Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry & research design: Choosing among five approaches*. SAGE Publications.
- Deldjoo, Y., Nazary, F., Ramisa, A., Mcauley, J., Pellegrini, G., Bellogin, A., & Di Noia, T. (2023). A review of modern fashion recommender systems. *ACM Computing Surveys*, 37(4), 111:1-111:37.
- Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019). On conversational agents in Information Systems Research: Analyzing the past to guide future work. In *Proceedings of the International Conference on Wirtschaftsinformatik*.
- Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). Towards understanding emotional intelligence for behavior change chatbots. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, United Kingdom.
- González-García, M., Moreno-Schneider, J., Ostendorff, M., & Rehm, G. (2023). Integration of a Semantic Storytelling Recommender System in Speech Assistants. *Proceedings of the Text2Story Workshop*, 5–11.
- Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 3–28.
- Haug, A., & Busch, J. (2016). Towards an ethical fashion framework. *Fashion Theory*, 20(3), 317–339.
- Hwangbo, H., Kim, Y. S., & Cha, K. J. (2018). Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*(28), 94–101.
- Kim, K. H., & Kim, E. Y. (2020). Fashion marketing trends in social media and sustainability in fashion management. *Journal of Business Research*, 117, 508–509.
- Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the role of Artificial Intelligence in personalized engagement marketing. *California Management Review*, 61(4), 135–155.
- Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt.
- Mintzberg, H., Raisinghani, D., & Théorêt, A. (1976). The structure of "unstructured" decision processes. *Administrative Science Quarterly*, 21(2), 246–275.
- Montenegro, J. L. Z., Da Costa, C. A., & Da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, 129, 56–67.
- Moon, K., Lee, J., & Lai, S. (2017). Key drivers of an agile, collaborative fast fashion supply chain. *Journal of Fashion Marketing and Management: An International Journal*, 21(3), 278–297.
- Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems*, 24(1), 485–501.
- Orlovsky, S. (1978). Decision-making with a fuzzy preference relation. *Fuzzy Sets and Systems*, 1(3), 717–723.
- Osburg, V., Davies, I., McLeay, F., & Yoganathan, V. (2021). Perspectives, opportunities and tensions in ethical and sustainable luxury: Introduction to the Thematic Symposium. *Journal of Business Ethics*, 169(2), 201–210.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70(1), 153–163.
- Pero, M., Arrigo, E., & Fionda-Douglas, A. (2020). Sustainability in fashion brands. *Sustainability*, 12(14), 5843–5846.
- Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human-Computer Interaction*, 37, 81–96.
- Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 1–29). Springer New York.
- Rima, S., Meriem, H., Najima, D., & Rachida, A. (2023). Toward a generative chatbot for an OER recommender system designed for the teaching community: General architecture and technical components. *The International Conference on Artificial Intelligence and Computer Vision*, 348–357.
- Sadovykh, V., & Sundaram, D. (2017). Decision making processes in online social networks: A comparative analysis of health and financial online social networks. In *Hawaii International Conference on System Sciences*.

- Sarker, S., Xiao, X., Beaulieu, T., & Lee, A. S. (2018). Learning from first-generation qualitative approaches in the IS discipline: An evolutionary view and some implications for authors and evaluators (Part 2/2). *Journal of the Association for Information Systems*, 19(9), 909–923.
- Shevchuk, N., & Oinas-Kukkonen, H. (2019). Influence on intention to adopt green IS: Boosting endogenous motivations with persuasive systems design. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.
- Statista. (2021). *Umsatzstärkste online-shops im fashion-segment in Deutschland im Jahr 2021*. <https://de.statista.com/prognosen/646166/top-online-shops-fashion-deutschland-ecommerce>
- Stöckigt, G., Schiebener, J., & Brand, M. (2018). Providing sustainability information in shopping situations contributes to sustainable decision making: An empirical study with choice-based conjoint analyses. *Journal of Retailing and Consumer Services*, 43, 188–199.
- Sutanto, J., Leidner, D. E., & Goutas, L. (2021). Influencing sustainable consumption through persuasive agent design. In *Proceedings of the International Conference on Information Systems (ICIS)*.
- Sysko-Romańczuk, S., Zaborek, P., Wróblewska, A., Dąbrowski, J., & Tkachuk, S. (2022). Data modalities, consumer attributes and recommendation performance in the fashion industry. *Electron Markets*, 32, 1279–1292.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Um, T., Kim, T., & Chung, N. (2020). How does an intelligence chatbot affect customers compared with self-service technology for sustainable services? *Sustainability*, 12(12).
- WCED. (1987). *Our common future*. Oxford University Press.
- Xiao, J., Yang, Z., Li, Z., & Chen, Z. (2022). A review of social roles in green consumer behaviour. *International Journal of Consumer Studies*, 1–38.
- Yadav, R., & Pathak, G. (2017). Determinants of consumers' green purchase behavior in a developing nation: applying and extending the theory of planned behavior. *Ecological Economics*, 134(1), 114–122.
- Zalando. (2023). *Zalando kündigt fashion assistant basierend auf ChatGPT an*. <https://corporate.zalando.com/de/technologie/zalando-kuendigt-fashion-assistant-basierend-auf-chatgpt>
- Zhang, S., & Balog, K. (2020). Evaluating conversational recommender systems via user simulation. *Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining*, 1512–1520.

Appendix II: Additional Publications

Appendix II provides an overview of additional publications and articles that are not part of this dissertation.

Journal Articles

Segneri, L.; Babina, N.; Hammerschmidt, T.; Fronzetti Caladon, A. & Gloor, P. (2024). *Too much Focus on your Health might be bad for your Health! – Reddit Users' Communication Style Predicts their Long COVID Likelihood*. PLOS One, <https://doi.org/10.1371/journal.pone.0308340>.

Passlack, N.; Hammerschmidt, T. & Posegga, O. (2025). *It was(n't) me: Vignette Experiments on Managers' Responsibility Attribution in AI-advised Decision-Making*. Behavior and Information Technology (BIT), <https://doi.org/10.1080/0144929X.2024.2431050>.

Passlack, N.; Hammerschmidt, T. & Posegga, O. (forthcomming 2025, 3rd revision round). *With Great Power Comes Great Responsibility: What Abilities Drive Humans to Interact with AI (Ir)responsibly?* Information Systems Frontiers (ISF).

Conference Proceedings

Passlack, N.; Heyder, T.; Klemm, F. & Posegga, O. (2023). *How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success (RIP)*. Proceedings of the 44th International Conference on Information Systems (ICIS), Hyderabad, India, <https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/15>.

Bauer, M.; Hammerschmidt, T. & Posegga, O. (2025). *Exploring AI Narratives: How and Why AI Is Portrayed Differently Among People*. Proceedings of the 31st Americas Conference on Information Systems (AMCIS), Montréal, Canada, https://aisel.aisnet.org/amcis2025/sig_aiaa/sig_aiaa/19/.

Passlack, N.; Hammerschmidt, T. & Posegga, O. (2025). *How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice*. Proceedings of the 31st Americas Conference on Information Systems (AMCIS), Montréal, Canada, https://aisel.aisnet.org/amcis2025/is_leader/is_leader/4/.

Book Chapters

Efinger, M.; Eich, X. L.; Heck, M.; Nguyen, D. P.; Özlü, H. I.; Heyder, T. & Gloor, P. (2023). *Netnography 2.0: A New Approach to Examine Crowds on Social Media*. Handbook of Social Computing.

Stolz, K.; Heyder, T.; Gloor, P. A. & Posegga, O. (2019). *Measuring Human-Animal Interaction with Smartwatches: An Initial Experiment*. Collaborative Innovation Networks. Springer International Publishing, Cham, 165-182.

Working Papers

Hammerschmidt, T.; Stolz, K.; Posegga, O. (in progress). *AI for All? Investigating Barriers to Creating Inclusive and Equitable AI Outcomes*.

Xiao, Z.; Chengli, S. & Hammerschmidt, T. (in progress). *Systematic Literature Review on Ethical AI*.

Maurya, H.; Dan, D. & Hammerschmidt T. (2023). *Is ChatGPT Inclusive Enough? An Experimental Design Approach Analyzing the Fairness of AI Agents' Responses*. Paper-A-Thon at International Conference on Information Systems (ICIS).