

Managing in the Age of AI: Understanding Responsibility Perceptions and Changes in Ability Requirements

Nina Katharina Passlack

University of Bamberg

Bamberg 2026

Managing in the Age of AI: Understanding Responsibility Perceptions and Changes in Ability Requirements

Dissertation zur Erlangung des akademischen Grades doctor rerum politicarum (Dr. rer. pol.) der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg vorgelegt von Nina Katharina Passlack.

Erstgutachter: Prof. Dr. Kai Fischbach

Zweitgutachter: Prof. Dr. Karl-Heinz Gerholz

Mitglied der Promotionskommission: Prof. Dr. Daniel Beimborn

Datum der Abgabe: 31.10.2025

Datum der Disputation: 10.02.2026

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



URN: urn:nbn:de:bvb:473-irb-117035x

DOI: https://doi.org/10.20378/irb-117035

Von der genannten Lizenzangabe ausgenommen sind folgende Bestandteile dieser Dissertation:

Paper II "A Review of How Different Views on Ethics Shape Perceptions of Morality and Responsibility within AI Transformation" (S. 115-201), Paper IV "How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice" (S. 204-214), Paper V "From Teachers to Cheerleaders? Teachers' Role Responsibilities in the Age of Generative AI" (S. 216-247), Paper VI "With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?" (S. 249-284), Paper VIII "Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice" (S. 296-326) und Paper X "How Organizations Ethically Respond to Decision Dilemmas in AI Deployment" (S. 339-406) stehen unter der CC-Lizenz CC BY.

Lizenzvertrag: Creative Commons Attribution 4.0
https://creativecommons.org/licenses/by/4.0/



Table of Contents

ACKNOWLEDGEMENTS ..... **III**

ZUSAMMENFASSUNG (GERMAN SUMMARY) ..... **V**

1 INTRODUCTION **1**

2 THEORETICAL AND CONCEPTUAL FOUNDATIONS **6**

2.1 Theoretical Framework [Paper I] 6

2.2 Understanding Responsibility Perceptions 10

2.3 Changes in Ability Requirements 13

3 RESEARCH METHODS **16**

3.1 Systematic Literature Reviews 17

3.1.1 Sampling and Data Collection *18*

3.1.2 Data Analysis *19*

3.1.3 Data Quality, Validity, and Reliability *20*

3.2 Semi-Structured Interviews & Open-Ended Questions 20

3.2.1 Sampling and Data Collection *21*

3.2.2 Data Analysis *21*

3.2.3 Data Quality, Validity, and Reliability *22*

3.3 Experimental Research 22

3.3.1 Sampling and Data Collection *23*

3.3.2 Data Analysis *24*

3.3.3 Data Quality, Validity, and Reliability *25*

4 KEY FINDINGS AND CONTRIBUTIONS **26**

4.1 How do Professionals Perceive Responsibilities in Human–AI Collaboration? [RQ1] 26

4.1.1 How Responsibilities in Human–AI Collaborations Can be Perceived [Paper II] *27*

4.1.2 How Professionals Attribute Responsibility in Human–AI Collaborations [Paper III] *28*

4.1.3 How Professionals' Responsibility Attribution is Influenced by Their Moral Identity [Paper IV] *29*

4.1.4 How Perceived Role Responsibilities May Change in Human–AI Collaborations [Paper V] *29*

4.2 What Abilities and Structural Conditions Enable Professionals to Engage in Responsible Human–AI Collaboration? [RQ2] 30

4.2.1 What Shapes AI Literacy for Responsible Human–AI Collaboration [Paper VI] *31*

4.2.2 How Human–AI Collaborations Transform Abilities and Ability Requirements [Paper VII] *32*

4.2.3 How AI Literacy Affects AI Use in Case of AI Errors [Paper VIII] *33*

4.2.4 How AI Literacy Impacts Perceptions of AI [Paper IX] *34*

4.2.5 How Organizations Can Foster Responsible Human–AI Collaboration [Paper X] *34*

5 DISCUSSION AND IMPLICATIONS **36**

5.1 Theoretical Implications 36
5.2 Practical Implications 40
5.3 Limitations & Future Research Avenues 43
6 CONCLUSION 47
REFERENCES 48
APPENDIX 62
Appendix I. Papers Included in the Cumulative Dissertation 62
Appendix II. Overview of All Works 407

Acknowledgements

I have always resonated with the idea that the journey is the destination—that growth lies not only in where we arrive, but in how we get there. And yet, I have come to believe that what truly shapes the meaning of that journey is the company we keep along the way. It is in shared moments, trusted companionship, and quiet acts of mutual support that any journey becomes truly worthwhile.

This dissertation was possible thanks to many, the thoughtful input of my PhD committee, the collaboration of my colleagues, and the steadfast support of my friends and family beyond its scope.

First and foremost, I owe heartfelt thanks to Prof. Dr. Kai Fischbach. Your encouragement set this journey in motion, and your unwavering trust and support carried me through. It has meant a lot to know I could always count on you. A heartfelt thanks to Prof. Dr. Karl–Heinz Gerholz. Working with you was inspiring and forward–looking, and your perspective and energy lifted this project in ways I will always value. I am equally grateful to Prof. Dr. Daniel Beimborn for completing my committee. Thank you for your constructive feedback and your purposeful guidance; you helped me bring this dissertation across the finish line.

A special note of gratitude goes to my allies from day one. Teresa Hammerschmidt (née Heyder), collaborating with you was inspiring, intense, and so much fun. Your positive attitude and optimism, true to the motto "Where there's a will, there's a way," kept our momentum high and our standards higher. Theresa Henn, thank you for your support, for many laughing fits, and for our Tuesday sports sessions together with Insa, which may or may not have ended at Kiosk Kunni. The time the three of us shared in Bamberg and our travels through India and Canada are among the memories I most treasure. You have profoundly shaped my PhD journey.

To the remaining (affiliate) members of the Department of Information Systems and Social Networks (SNA): thank you for the everyday support that makes research possible. Dr. Oliver Posegga, thank you for your cooperation over the years; I am grateful for what I learned through our discussions. Dr. Kathrin Eismann and Dr. Diana Fischer, thank you for the supportive calls and your mentorship over coffee. Falco Klemm: thanks for jumping in on every department and IT question. Scott Cooper, thank you for making learning English grammar unexpectedly fun. The chair's student assistants (HiWis), thank you for your help on our projects. And a very special thanks to Gudrun Stilkerich. Your warm personality and your help with all administrative things were invaluable.

I am also deeply grateful to the Wirtschaftspädagogik (WiPäd) Bamberg team at the Professorship for Business and Human Resource Education for our collaborations across diverse projects and the business trips we shared. Working with you was a joy: Dr. Christina Bader, Shaleen Beil, Anna Benning, Lisa Gabriel, Christian Gärtner, Silvia Leicht, Philipp Schlottmann, Hui Tang, Dr. Anne Wagner, Judith Wittig, the chair's HiWis, and the entire department. Thank you for your cooperation and for the many ways you kept things moving.

My sincere thanks to all my "external" coauthors outside the University of Bamberg for stimulating discussions, and each of your personal contributions to my work and PhD years: Dr. Dirk Leffrang and Prof. Dr. Oliver Müller (University of Paderborn); Dr. Katharina Stolz (University of Stuttgart); and

Alina Hafner (Technical University of Munich). Your individual contributions and our interdisciplinarity broadened my perspective and strengthened this dissertation.

As balance matters, thank you to my friends who ensured there was life beyond the PhD. I am grateful to my ERBA Squad, which began with Julian Frey, Ferdinand and Elisabeth Mittermaier, and Yannick Hildebrand, and has grown into a family of more than twenty members. Smashing conference nights in India and Hawaii, and sharing stories and feelings over a cold drink, have always been grounding. Thanks also to David Beck. You always shared your thoughts and challenged my ideas; I am certain our coffee sessions strengthened my work. And thank you to the Bamberg Choir "Mädelsabend" for providing a welcome respite—shameless plug over. A heartfelt thank you also goes to my friends Theresa Hatz, Anna Sonntag, Raphael Helmreich, Alexander Hochstein, and Lena Kühn—just to name a few who have brightened my life for almost (some more than) twenty years now. To the many others, thank you. You know who you are.

Last but not least, I extend my deepest gratitude to my family, whose love has supported me every step of the journey. My parents, Anna–Maria and Rainer, my sister Julia, and, in loving memory, my grandparents Doris and Karl–Werner, who from an early age encouraged the confidence in me to trust myself and my abilities, the joy of engaging with new topics and others, and the appreciation for whatever life brings. Also, thanks to my brother-in-law Aron for being my go-to reality check whenever I needed to ensure that my research still made sense to practitioners. One of my biggest expressions of gratitude goes to Robin. Thank you for your love and your endless support—even when that meant packing a laptop on every holiday, even our road–bike tours, so I could write during breaks by the lake.

To all those who walked beside me, thank you for having made my PhD journey truly worthwhile.

Nina Passlack

Zusammenfassung (German Summary)

Mit der zunehmenden Integration Künstlicher Intelligenz (KI) in Organisationen stellt sich für das Management mit Dringlichkeit die Frage nach den Implikationen und Konsequenzen der Zusammenarbeit zwischen Menschen und KI. Neben strategischen Top-down-Initiativen zur Effektivitätssteigerung durch KI-Einsatz nutzen Beschäftigte zunehmend niedrigschwellige Applikationen wie Chatbots (zum Beispiel ChatGPT, Claude oder GitHub Copilot) zur Unterstützung individueller Aufgaben und tragen somit zur Demokratisierung von KI bei (Krakowski, 2025). Mensch-KI-Kollaborationen sind ein zentrales Forschungsthema der Wirtschaftsinformatik als soziotechnischer Disziplin (Sarker et al., 2019), da sie Arbeitsprozesse, Organisationsstrukturen und professionelle Rollen tiefgreifend verändern können (Angelopoulos et al., 2023; Boos et al., 2013; Grodal et al., 2023; Krakowski, 2025).

Im Unterschied zu klassischen Technologien lernen KI-Systeme, etwa Deep-Learning-Modelle, virtuelle Assistenten oder Verfahren der Sprachverarbeitung (Natural Language Processing) (Benbya et al., 2021), aus Daten und generieren probabilistische Ergebnisse, die nicht immer kausal nachvollziehbar sind (e.g., Floridi et al., 2021). Daraus folgt, dass Beschäftigte weiterhin in der Verantwortung stehen, Ergebnisse kritisch zu prüfen, Domänenkontext einzubringen und Verantwortung für die Konsequenzen der Mensch-KI-Kollaboration zu übernehmen (Martin, 2019). Zugleich konfrontieren Intransparenz und Selbstlernfähigkeit der KI-Systeme Organisationen mit Herausforderungen, denn Modellergebnisse sind oft nur begrenzt nachvollziehbar (Berente et al., 2021) und Nebenfolgen wie Diskriminierung (e.g., Verma, 2022) oder Fehlerfortpflanzungen aufgrund fehlerhafter Datengrundlagen (Lebovitz et al., 2021) lassen sich schwer identifizieren und wirksam verhindern.

Diese Dissertation nimmt eine relationale Perspektive ein und versteht die Funktionalitäten und Handlungsmöglichkeiten von KI als etwas, das dynamisch in der Nutzung und Interaktion entsteht (Hinds & Von Krogh, 2024). Sie ergänzt damit den verbreiteten Fokus auf Effizienzgewinne um eine systematische Betrachtung unbeabsichtigter Effekte, der Verlust kritischer Reflexionsfähigkeit (Passlack, Gerholz, et al., 2025) oder wahrgenommene Gefährdung der beruflichen Identität aufgrund des Risikos, dass Beschäftigte zu bloßen Übermittlern von Modellergebnissen werden, statt eigenständig Wert zu schaffen (Harder Fischer et al., 2024; Strich et al., 2021). Daraus leitet sich die Notwendigkeit einer verantwortungsorientierten – und nicht primär technikzentrierten – Gestaltung von Mensch-KI-Kollaborationen ab. Im Zentrum stehen zwei leitende Forschungsfragen: Wie kann Verantwortung in der Zusammenarbeit mit KI wahrgenommen werden? (RQ1) Und welche Fähigkeiten sowie strukturellen Rahmenbedingungen ermöglichen verantwortungsvolle Zusammenarbeit mit KI? (RQ2)

Zur Beantwortung dieser Fragen umfasst die Arbeit zehn Beiträge, die nachfolgend als Paper referenziert werden. Die theoretische Fundierung für beide Forschungsfragen bildet **Paper I,** indem es soziotechnische Systemtheorie mit dem soziomateriellen Ansatz verbindet und damit den theoretischen Rahmen bietet. Hinsichtlich der ersten Forschungsfrage (RQ1) beinhaltet die Dissertation vier Papiere: **Paper II** untersucht auf Basis einer Literaturrecherche, wie Verantwortung von KI-Systemen wahrgenommen wird. **Paper III** liefert empirische Evidenz durch online Experimente zu Verantwortungsattribution in KI-gestützten Entscheidungsprozessen, während **Paper IV** den Einfluss

moralischer Identität auf eben diese Verantwortungsattribution analysiert. **Paper V** beleuchtet erste Verschiebungen professioneller Rollenverständnisse im Bildungskontext durch Interviews.

Zur Beantwortung der zweiten Forschungsfrage (RQ2) tragen fünf Papiere bei: **Paper VI** entwickelt das Konzept einer verantwortungsorientierten KI-Literacy durch eine systematische Literaturanalyse mit anschließenden Experteninterviews. **Paper VII** untersucht mit Hilfe von Interviews, wie sich Fähigkeiten und Fähigkeitsanforderungen durch KI-Einsatz verändern. **Paper VIII** analysiert mit Hilfe von Online-Experimenten, welchen Einfluss KI-Literacy auf die Nutzung von KI-Systemen hat, und **Paper IX** untersucht Narrative und dabei den Einfluss von KI-Literacy auf die Wahrnehmung von KI. **Paper X** schließlich formuliert erste Empfehlungen, wie Organisationen verantwortungsvolle Mensch-KI-Kollaborationen fördern und Entscheidungsdilemma im Umgang mit KI adressieren können.

Methodisch stützt sich die Dissertation somit auf eine Kombination aus systematischen Literaturanalysen, semi-strukturierten Interviews und experimentellen Studien. Dadurch leistet sie sowohl einen Beitrag zur theoretischen Weiterentwicklung der Wirtschaftsinformatik als auch konkrete, anwendungsnahe Implikationen für Organisationen, die die Potenziale von KI ausschöpfen möchten, ohne deren Risiken zu vernachlässigen. Insgesamt zeigt die Arbeit, wie Verantwortungszuschreibungen in Mensch-KI-Kollaborationen entstehen können, welche Kompetenzen und Rahmenbedingungen zu ihrer verantwortungsvollen Ausgestaltung erforderlich wären und wie diese Erkenntnisse in Forschung und Praxis fruchtbar gemacht werden können.

Theoretisch leistet die vorliegende Dissertation, über die Kernergebnisse der einzelnen Papiere hinaus, präzisierbare Implikationen für die Wirtschaftsinformatik. Erstens zeigt die Arbeit wie eine soziomaterielle Perspektive hinsichtlich möglicher Implikationen von Mensch-KI-Kollaborationen in Organisationen systematisch Transparenz bietet. Zweitens legt die Unterscheidung zwischen einem subjektiven Verantwortungsgefühl und formaler Verantwortlichkeit bislang verdeckte Spannungslagen offen. Drittens erfordert eine verantwortungsorientierte Mensch-KI-Zusammenarbeit eine erweiterte Konzeption von KI-Literacy, die kontextuelle, individuelle sowie technikzentrierte Einflussfaktoren ausdrücklich einbezieht. Viertens macht die Analyse von informellen Lernprozessen Qualifikationslücken sichtbar, die aus der Spannung zwischen notwendiger Weiterqualifizierung und gleichzeitigem Rückgang bestehender Kompetenzen entstehen. Fünftens kann KI-Literacy zwar das Funktionsverständnis und die Handlungssicherheit im Umgang mit KI erhöhen, paradoxerweise jedoch das Vertrauen in KI verringern.

Aus den Beiträgen ergeben sich auch praxisrelevante Implikationen. Zum einen zeigt die Arbeit, dass Verantwortungslücken in Mensch-KI-Entscheidungen nicht primär technisch gelöst werden können, sondern eine Führungsaufgabe sind. Führungskräfte sollten klare Zuständigkeiten und transparente Begründungsstandards verankern, damit Verantwortung sichtbar bleibt und nicht an die KI delegiert wird. Zudem ist KI-Literacy als ein Werkzeugkasten zu verstehen, der je nach Bereich des KI-Einsatzes, individuellen Faktoren und je nach Verantwortlichkeiten unterschiedlich bestückt werden muss. Qualifizierung sollte zudem auf Augmentation statt Substitution ausgerichtet sein. Abschließend skizziert die Dissertation fünf mögliche Ansatzpunkte für zukünftige Forschung. Damit werden konkrete Perspektiven für die Weiterentwicklung des Forschungsfeldes aufgezeigt. Sie versteht sich also als Beitrag, der künftige wissenschaftliche Arbeiten zum Thema verantwortungsvolle Mensch-KI-Kollaboration anregt und unterstützt.

1 Introduction

Artificial Intelligence (AI) has moved to the center of organizational practice. Alongside top-down efforts to streamline workflows, such as managerial decision-making, employees increasingly adopt easily accessible tools such as ChatGPT and Claude to support individual tasks (Krakowski, 2025). Given AI's profound impact on work processes, organizations, and individual roles (Angelopoulos et al., 2023; Boos et al., 2013; Grodal et al., 2023; Krakowski, 2025), human–AI collaboration has become a core concern in information systems (IS) as a sociotechnical discipline (Sarker et al., 2019). As the term "artificial intelligence"—spanning technologies such as deep learning, neural networks, virtual assistants, robotics, and natural language processing (Benbya et al., 2021)—suggests, these technologies are perceived as exhibiting forms of machine intelligence that surpass the capabilities of previous ones as they learn from data and act not only based on explicit instructions (Munoko et al., 2020, p. 2).[1] AI's outputs are, thus, probabilistic, based on correlations and not necessarily clear cause-and-effect relationships (Floridi et al., 2021). This keeps professionals firmly "in the loop" for responsible use (Martin, 2019) to check plausibility and supply domain context and to assume accountability. Accordingly, this dissertation adopts a relational view of human–AI collaboration, meaning that AI's functionality and affordances emerge dynamically through use and collaboration (Hinds & Von Krogh, 2024) and that humans and AI achieve superior results compared to either operating alone (Kamar, 2016).

Prevailing narratives emphasize AI-enabled performance and task efficiency, yet recent work foregrounds potential negative side effects and unintended consequences (Papagiannidis et al., 2025; Retkowsky et al., 2024). A narrow focus on perfect outcomes of human–AI collaboration obscures unintended consequences—sometimes referred to as the "dark side of AI" (Mikalef et al., 2022)—that affect individuals, organizations, or society at large (Benbya et al., 2021; Passlack et al., 2024; Trier et al., 2023). Unintended consequences of human–AI collaboration may include issues such as discrimination and exclusion of minorities (Verma, 2022), behavior lacking in self-reflection (Passlack, Gerholz, et al., 2025), downstream harms when ground truth is flawed (Lebovitz et al., 2021), and perceived threats to professional identity from the risk that workers become mere conveyors of model outputs rather than producers of value (Harder Fischer et al., 2024; Strich et al., 2021). To secure long-term success for organizations, this dissertation argues for a deliberate approach that emphasizes *responsibility*, aiming for "fair, and value-based" human–AI collaboration "beyond the legal minimum" (Trier et al., 2023, p. 463).

Responsibility in human–AI collaboration, however, remains a subject of discussion. Humans are ultimately responsible for moral considerations (Constantinescu et al., 2021; Dyrkolbotn et al., 2018) despite that AI's increasing self-learning ability (Berente et al., 2021) implies that these technologies are no longer consistently under human control (Baird & Maruping, 2021), especially when model architectures are opaque or non-transparent (Rowe et al., 2024; Susser, 2019). Given this opacity, it is

[1] When discussing meta-level implications that are not tied to a particular technology, this dissertation uses the broader term "AI." Where findings were related to a specific technology such as "generative AI," it is named explicitly.

also unrealistic to expect that designers can always foresee every action of AI systems (McDonald & Pan, 2020), which shifts practical responsibility to the professionals who collaborate with and deploy these systems (Heyder et al., 2023). Meanwhile, as AI exhibits more humanlike behaviors and capabilities (Chandra et al., 2022; Li, 2021), people may treat it less as a tool and more as "a genuine team member" (O'Neill et al., 2022, p. 907), potentially undermining established structures of responsibilities within organizations (Smith & Vickers, 2021) and leading to less critical reflection among users (Passlack et al. 2025). While legal responsibility is comparatively well-covered in human–AI collaborations, moral responsibility remains underexplored and warrants further investigation (Rowe et al., 2024). In particular, since "responsible AI" is often framed as compliance with broader societal norms, scholarship has paid far less attention to the individual's role within organizational practice (Papagiannidis et al., 2025). In this regard, responsibility gaps on an individual level among professionals need to be made transparent and be better understood, aiming for responsible human–AI collaboration.

The first research objective (***RQ1***) this dissertation addresses is thus:

RQ1: How do professionals perceive responsibilities in human–AI collaboration?

Furthermore, following Matthias (2004), the attribution of responsibility for actions or outcomes in human–AI collaboration presupposes that individuals possess sufficient comprehension of the system; without such understanding, responsibility cannot meaningfully be assumed. Yet, improved usability has opened AI technologies to a much wider audience, enabling people to use them effectively regardless of their technical background (Maedche et al., 2019). Meanwhile, the complexity of these technologies continues to grow, posing significant challenges to users' ability to understand the processes behind AI and its implications (Berente et al., 2021). This suggests that responsible human–AI collaboration calls for novel abilities, competencies, and skills (Berente et al., 2021; Y. K. Dwivedi et al., 2023; Rana et al., 2022). Thus far, existing conceptualizations of AI literacy (e.g., Cetindamar et al., 2022; Pinski & Benlian, 2023; Wang et al., 2022)—defined 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 & Magerko, 2020, p. 598)—focus mainly on economic- and performance-oriented outcomes, rather than considerations of responsibility. Responding to calls for "harnessing benefits while addressing associated risks and concerns" (Carroll et al., 2024, p. 854), this dissertation adopts a responsibility-oriented lens that explicitly addresses unintended outcomes and decision dilemmas in search of required abilities and structural conditions that enable responsible human–AI collaboration.

The second research question (***RQ2***) guiding this dissertation is thus:

RQ2: What abilities and structural conditions enable professionals to engage in responsible human–AI collaboration?

This dissertation comprises ten papers (listed in Table 1) that respond to the two research questions. There are three published conference papers and seven journal articles, four of which are published, one which is a Registered Report with in-principle acceptance (currently in the second round of Stage-2

review),[2] and two that are under review at the time of this writing. Published in outlets ranked from B to A in the VHB Rating 2024,[3] all papers underwent a double-blind peer review process.

The ten papers contribute to this dissertation's research questions and IS research in several ways. **Paper I** serves as a theoretical foundation and lens of this dissertation that provides a deeper understanding of phenomena emerging in human–AI collaborations by integrating sociotechnical systems theory with the lens of sociomateriality (**Paper I**). Four papers contribute to the first research objective (***RQ1***) by investigating perceptions of responsibility in human–AI collaborations. Building upon a theoretical inquiry into how individuals may perceive the notion of responsibility with respect to AI systems (**Paper II**), the dissertation offers empirical evidence on responsibility attribution within AI-supported decision-making contexts (**Paper III**) and investigates how such attributions are shaped by individuals' moral identity (**Paper IV**). In addition, it presents preliminary findings on how human–AI collaborations may influence perceptions of role-specific responsibilities (**Paper V**).

Five papers contribute to the second research objective (***RQ2***) by investigating abilities and structural conditions that enable professionals to engage in responsible human–AI collaboration. Based on interviews, this dissertation investigates how the concept of AI literacy needs to be developed for responsible collaborations with AI (**Paper VI**). The dissertation then explores how the ability requirements of individuals change in response to their AI use (**Paper VII**), and offers initial insights into how AI literacy may influence both the use of AI advice (**Paper VIII**) and individual perceptions of AI (**Paper IX**). Finally, **Paper X** presents preliminary recommendations for how organizations can cultivate responsible human–AI collaboration among professionals by responding to decision dilemmas that may occur in AI deployment.

As Figure 1 outlines, the remainder of this dissertation is structured as follows. Next, the theoretical and conceptual foundations are presented. This includes **Paper I**, which builds the theoretical framework and guides the development of and sets the stage for the presentation of the remaining papers in section 2. The details on the methods used to find answers to the dissertation's two research questions follow in section 3, where the dissertation draws on different methods related to systematic literature reviews, semi-structured interviews, and experimental studies. The key findings of this dissertation, organized according to the two research questions, are presented in section 4. Section 5 discusses the results and their broader implications for research and practice, outlines the limitations of this dissertation, and presents five avenues for future research. Finally, there is a short conclusion in section 6.

[2] Of note, our work is the first in–principle acceptance for a Registered Report in the journal BISE (https://bise-journal.com/?page_id=2258).

[3] VHB Rating 2024: https://www.vhbonline.org/services/vhb–rating–2024.

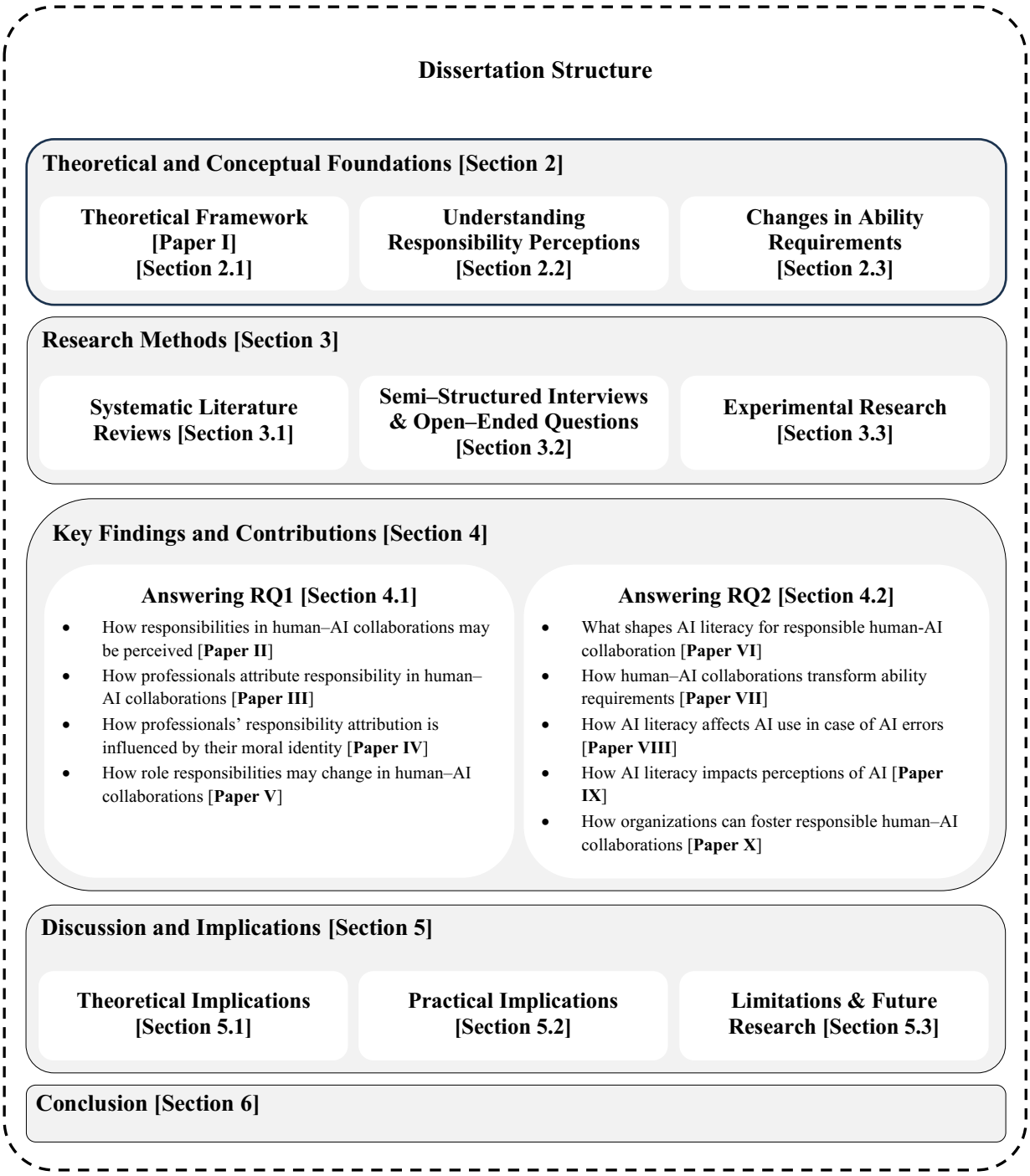


Figure 1. Dissertation Structure

Table 1. Classification of Papers Contributing to this Dissertation

Paper	Contribution Type [J vs C]*	Contributors	Title	Research objective	Publication organ; VHB Rating 2024	Status
I	Theory development based on a systematic literature review [J]	Heyer, **Passlack**, Posegga	Ethical management of human–AI interaction: Theory development review	Understanding the ethical management of human–AI collaborations	The Journal of Strategic Information Systems (JSIS); A-ranked (VHB JOURQUAL-4)	Published (2023)
II	Theoretical-driven propositions based on a systematic literature review [J]	Hammerschmidt, Hafner, Stolz, **Passlack**, Posegga, Gerholz	A review of how different views on ethics shape perceptions of morality and responsibility within AI transformation	How responsibilities in human–AI collaborations can be perceived	Information Systems Frontiers (ISF), B-ranked (VHB JOURQUAL-4)	Published (2025)
III	Empirical evidence based on online experiments [J]	**Passlack**, Hammerschmidt, Posegga	It was(n't) me: Vignette experiments on managers' responsibility attribution in AI-advised decision-making	How professionals attribute responsibility in human–AI collaborations	Behaviour and Information Technology (BIT); B-ranked (VHB JOURQUAL-4)	Published (2024)
IV	Empirical evidence based on online experiments [C]	**Passlack**, Hammerschmidt, Posegga	How leaders' morality influences their responsibility perceptions when relying on algorithmic advice	How professionals' responsibility attribution is influenced by their moral identity	America's Conference on Information Systems (AMCIS); C-ranked (VHB JOURQUAL-4)	Published (2025)
V	Empirical evidence & propositions based on semi-structured interviews [J]	**Passlack**	From teachers to cheerleaders? Teachers' role responsibilities in the age of generative AI	How perceived role responsibilities may change in human–AI collaborations	Previously passed the 1st revision round at Journal of Information Technology (JIT); A-ranked (VHB JOURQUAL-4)	Working Paper (unpublished)
VI	Theoretical-driven conceptualization, empirical evidence & propositions [J]	**Passlack**, Hammerschmidt, Posegga	With Great Power Comes Great Responsibility: What shapes AI literacy for responsible interactions of knowledge workers with AI?	What shapes AI literacy for responsible human–AI collaboration	Information Systems Frontiers (ISF), B-ranked (VHB JOURQUAL-4)	Published (2025)
VII	Empirical evidence & propositions based on semi-structured interviews [C]	**Passlack**, Gerholz, Schlottmann	Reskilling me softly: Perceived changes in students' skilling when using genAI in academic research projects	How human–AI collaborations transform abilities and ability requirements	Hawaii International Conference on System Science (HICSS); B-ranked (VHB JOURQUAL-4)	Published (2025)
VIII	Empirical evidence & propositions based on online experiments [J]	Leffrang, **Passlack**, Müller, Posegga	Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice	How AI literacy affects AI use in case of AI errors	Business & Information Systems Engineering (BISE); B-ranked (VHB JOURQUAL-4)	Published (2026)
IX	Empirical evidence & propositions based on online questionnaires [C]	Hammerschmidt, **Passlack**, Stolz, Posegga	AI narratives: What can they tell us about individuals' AI literacy and emotional attitude toward AI assistants?	How AI literacy impacts perceptions of AI	Hawaii International Conference on System Science (HICSS); B-ranked (VHB JOURQUAL-4)	Published (2024)
X	Empirical evidence & propositions based on semi-structured interviews [J]	Hammerschmidt, **Passlack**, Posegga	How organizations ethically respond to decision dilemmas in AI deployment	How organizations can foster responsible human–AI collaborations	Previously progressed to a 2nd revision round at The Journal of Strategic Information Systems (JSIS); A-ranked (VHB JOURQUAL-4)	Working Paper (unpublished)

* J = journal article; C = conference article

2 Theoretical and Conceptual Foundations

The following sections establish this dissertation's theoretical background. Section 2.1 presents the overarching framework that anchors the remaining work. Section 2.2 provides the foundation for RQ1 by defining responsibility, its relation to morality, ethics, and agency, and by unpacking the psychological mechanisms of responsibility attribution. Section 2.3 then develops the background for RQ2 with respect to changes in ability requirements, articulating the conceptual dimensions of AI literacy, and outlining the core dimensions of reskilling and deskilling.

2.1 Theoretical Framework [Paper I]

Paper I establishes the theoretical groundwork for this dissertation by introducing the core conceptual framework that shapes the academic lens of the subsequent papers. It aids in systematically examining how technologies and humans mutually shape each other and are entangled in organizational life. Based on a systematic literature review, we engaged in theorizing to better understand the management of human–AI collaborations with an ethical perspective, seeking responsible outcomes. By relying on the sociotechnical systems framework from Leonardi (2012) and extending it with more nuanced reflections on the notions of sociomateriality (*inseparability*, *performativity*, *relationality*, *practices*, and *materiality*) as brought forward by Jones (2014), it proposes the integrated framework adapted in Figure 2. This integrated framework helps in better understanding human–AI collaborations in a systematic way.

In general, human–AI collaboration can be categorized according to different types (*automation*, *augmentation*, *assemblage*). Recent literature argues that emerging human–AI collaboration is shifting from pure *automation*, where technologies assume tasks previously undertaken by humans, toward *augmentation*, meaning that human abilities can be enhanced through AI and for which some human weaknesses can be compensated (Daily et al., 2017; Raisamo et al., 2019). In some cases, we can speak of *assemblage*, which goes one step further to suggest that humans and AI become, more or less, a single integrated unit in solving tasks (Rai et al., 2019). In so doing, they achieve outcomes that would not have been possible separately (O'Neill et al., 2022). We take the stance that we cannot classify specific human–AI collaboration as *one* of those types (**Paper I**). Rather, we need to adopt a process view, as "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" (Heyder et al., 2023, p. 9). Therefore, this dissertation uses the term human–AI collaboration to mean the interaction and partnership between humans and AI-based systems to achieve shared objectives in a way that neither could effectively accomplish without the other (Kamar, 2016).

Conceptualizing human–AI collaboration in organizations as sociotechnical systems and analyzing them through the lens of sociomateriality provides a comprehensive framework that enhances its understanding in several ways. The sociotechnical perspective can inform the management of human–AI collaborations by describing how AI technologies (as a material entity) are designed, adopted, and made meaningful through social phenomena when social and material entities are entangled (Leonardi, 2012). This entanglement between AI (as a material entity) and humans (as social entities) can be described as sociomaterial practices in which outcomes emerge from ongoing interweaving of humans and technology in practice (Leonardi, 2012). The sociomaterial practices influence and are influenced

by the social subsystem, which is made up of social objects (e.g., humans, organizational norms, culture, and role identities) and material objects (e.g., technologies, regulations) (Leonardi, 2012).

Paper I contributes to previous research by offering insights into how the lens of sociomateriality (as brought forward by Jones (2014)) can be applied systematically in sociotechnical systems, such as organizations, to better explain how material entities (such as technology) and social entities (such as managerial processes, culture, and human factors) are interdependent and interwoven—sometimes referred to as "imbrication" (Leonardi, 2012)—and what this means for ethical management of human–AI collaboration. Further, it explains how human moral character, AI's codified obligations, and an organization's social frameworks may shape one another.

Jones (2014) summarized the discourse on sociomateriality and introduced the notions of sociomateriality (*inseparability, relationality, performativity, materiality,* and *practices*) that help in understanding such interdependencies. These notions serve as a theoretical lens to explain and better understand implications of human–AI collaborations and how they are interrelated with management practices. Based on **Paper I**, they can be summarized as follows.

- *Inseparability* refers to the extent to which the social and material entities are perceived as distinct or intertwined. **Paper I** suggests that stronger forms of inseparability may contribute to responsibility gaps, as it becomes increasingly difficult to discern whether certain (un)intended outcomes stemmed from the influence of the social or the material entity. For example, in the context of autonomous driving, human drivers remain formally responsible for the vehicle's operation, yet AI systems continuously monitor the environment and intervene in critical situations, such as when an obstacle appears on the road. Although humans possess the final decision-making authority, their limited situational awareness at the moment of AI intervention may impair their ability to act meaningfully (Noy et al., 2018), thereby blurring the attribution of responsibility (Rowe et al., 2024). **Paper III** and **Paper IV** explore responsibility attribution in human–AI collaborations, where AI advice supports decision-making, in greater depth.
- *Performativity* concerns how human actors (the social entity) construct their reality, which is influenced by certain boundaries and limitations, often shaped by the broader social subsystem (Cecez-Kecmanovic et al., 2014). For instance, media can explain the tendency of humans to increasingly perceive AI as a "teammate" (O'Neill et al., 2022), which may influence individuals' perceptions of responsibilities and their reliance on AI advice. This is examined more deeply in **Paper II**, which explores varying perceptions of responsibility and moral agency of AI, and in **Paper IX**, which investigates how individual knowledge about and experience with AI influences perception of AI.
- *Relationality* addresses the ways in which the sociomaterial entanglement between the social and material entities influences both. This includes effects on human agency, such as losing certain competencies when employees no longer need to complete tasks alone but are supported by AI-based technologies (Mirbabaie et al., 2022). This may impact in particular new employees who have never learned to perform these tasks without AI support. Moreover, AI-enabled application systems may marginalize consultants' participation in decision-making, with consequential effects on their professional identity and role self-perception (Mayer et al., 2020). These dynamics are further explored in **Paper VI** and **Paper VII,** which

examine how characteristics of the social entity, namely individual abilities, are shaped through collaboration with AI as the material entity. **Paper V** investigates how AI usage leads to potential shifts in professional role responsibilities.

- *Practices* refer to how human–AI collaborations are embedded within organizational processes (social subsystems), and how the routines and characteristics of the material entity, alongside characteristics of the social entity, influence these work processes. For instance, algorithms assist Uber in identifying optimized routes; however, not all drivers adhere to these suggestions (Pregenzer et al., 2021). This illustrates how individual characteristics influence the ways in which people respond to or act through a material entity. This suggests that the outcome of the sociomaterial practices can affect humans' future practices as highlighted in **Paper VIII**, which investigates how the performance of AI systems affects individual advice-taking behavior and how this is affected by individual AI literacy. **Paper X** further explores how organizations respond to decision dilemmas that may arise concerning human–AI collaborations.
- *Materiality* focuses on the material entity and how objects of the social subsystem and the material entity's physical and non-physical objects are arranged. It refers to rather hardwired attributes of technological objects whose properties are constant across times and settings (Leonardi, 2012). For instance, the output, regardless of whether it is ethical, is strongly connected to the data derived from the social subsystem. Given this dissertation's human-centered focus on the social entity, further in-depth analysis of the notion of materiality lies beyond the scope of this work.

Figure 2 illustrates how the papers are embedded in the dissertation's theoretical framework.

Paper I, in synthesizing the discourse on sociomateriality (among others, Cecez-Kecmanovic et al., 2014; Jones, 2014; Orlikowski, 2007, 2010), suggests that these five notions of sociomateriality influence each other in cycles (suggesting "The Sociomateriality Cycles"), highlighting that understanding the implications of human–AI collaboration implications on individuals, organizations, and management is complex. *Performativity* refers to individuals' initial perceptions and expectations of AI prior to engaging in direct collaboration, which is likely influenced by the social subsystem through norms and culture. In contrast, *inseparability* captures the moment of entanglement, the point at which a person and AI as the material entity interact directly. This interaction informs *relationality*, as the collaboration influences a user's characteristics, capabilities, and perceived agency, thus, making it a consequence of the sociomaterial entanglement. Furthermore, the entanglement between human (as a social entity) and AI (as a material entity) leads to shifts in organizational routines and practices, aligning with the notion of *practices*. These changes in practices can, in turn, reshape future *performativity*. This illustrates how these dimensions interact in a cyclical and mutually reinforcing manner.

Paper I further contributes to IS research by suggesting seven avenues for future research. It identifies four effects in human–AI collaboration, one of which is the responsibility attribution effect, suggesting that individuals may attribute responsibility to AI and calling for future investigations (*Avenue 1*). It proposes using sociomaterial notions as a theoretical foundation to better understand key considerations in human–AI collaboration (*Avenue 2)* and calls for more human-centered approaches that place individuals at the core of investigations (*Avenue 3)*. It also recommends adopting the Next-Generation

Theorizing approach of Burton-Jones et al. (2021), which encourages researchers to see theory as a dynamic and ongoing intellectual conversation and urges scholars to learn with and from practitioners in order to better understand and theorize real-world challenges, moving theory-building beyond the desk and into dialogue with practice (*Avenue 4)*. Moreover, it advocates for integrating ethical considerations at a strategic level within management (*Avenue 5*) and emphasizes the need for management to make the workforce literate for responsible human–AI collaboration (*Avenue 6)*. Finally, it encourages the integration of diverse stakeholder perspectives to facilitate discussions on how to achieve responsible human–AI collaboration within organizations (*Avenue 7*).

These seven avenues have informed the development of this dissertation by shaping its theoretical foundation, methodological orientation, and overarching research objectives: The remaining nine papers are viewed as a focused exploration of or "zoom-ins" into the different notions of sociomateriality (*Avenue 2*), while the methodological approach is guided by the principles of Next-Generation Theorizing (*Avenue 4*). The dissertation adopts a human-centered perspective (*Avenue 3*), integrating different stakeholder perspectives such as those of managers, professionals, educators, and developers (*Avenue 7*), and underscores the ethical dimensions of management at a strategic level, particularly in fostering responsible, long-term outcomes (*Avenue 5*). Finally, these avenues also underpin the dissertation's two primary research objectives: develop a deeper understanding of how responsibility is perceived in the context of human–AI collaboration (RQ1; *Avenue 1*); and investigate the ability requirements essential for preparing professionals to collaborate with AI responsibly (RQ2; *Avenue 6*).

The following pages present the theoretical background for the two research questions: first by delineating core concepts and relevant scholarly discourse aimed at deepening the understanding of responsibility perceptions in human–AI collaborations (section 2.2, related to **RQ1**); and then by outlining concepts relevant to explaining what abilities and structural conditions enable professionals to engage in responsible human–AI collaboration (section 2.3, related to **RQ2**).

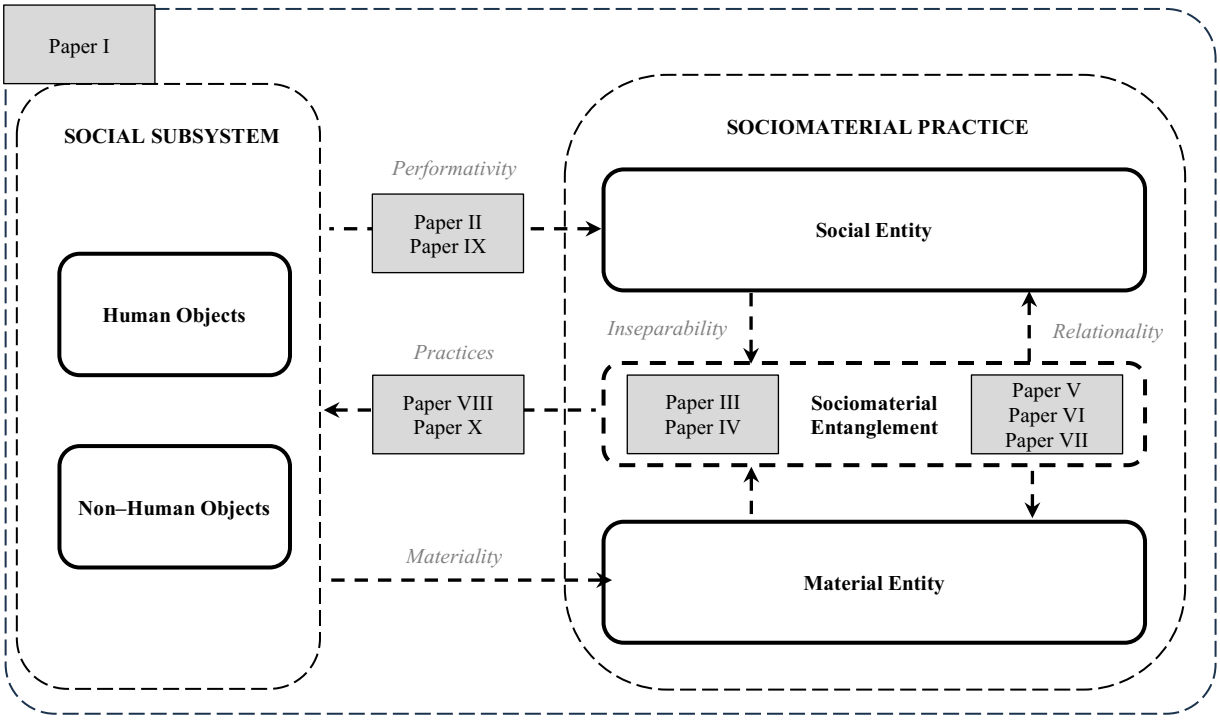


Figure 2. Embedding of the Papers into this Dissertation's Theoretical Framework (Adapted from Paper I)

2.2 Understanding Responsibility Perceptions

To address the first research objective (RQ1)—examining perceptions of responsibility in human–AI collaboration—this section clarifies the core constructs and outlines their relationships. It begins by defining *responsibility* as encompassing both forward-looking obligations (duties) and backward-looking evaluations (blame or credit for outcomes) (Van De Poel, 2011). I then define *responsibility attribution*—the psychological judgment by which observers assign responsibility to themselves or others for specific consequences—and explain *attribution theory*, which describes how people infer causality and control when allocating responsibility (Weiner, 1985). Because responsibility judgments rest on normative foundations, I also differentiate *morality*—the shared expectations about right and wrong that guide conduct (Barnes, 1999)—from *ethics*, the systematic reflection on, and justification of, those moral values (Churchill, 1999). Finally, I explicate *moral agency*, understood as an actor's perceived capacity for intention and control, which is treated here as a prerequisite for meaningful attributions of responsibility (Constantinescu et al., 2021). The subsections that follow define each construct in detail.

Responsibility is a term that commonly refers to a human's duty or obligation to perform a task, defined by roles or grounded in ethics through norms; thus, it is not always formal (Cañas, 2022). Responsibilities are often tied to particular professional roles, which this dissertation refers to as *professional role responsibilities*. These are often shaped by shared expectations regarding how individuals in a given position are expected to act (Stets et al., 2020). While sometimes used as a synonym for responsibility, *accountability*, in contrast, is linked to external expectations or authority, representing the obligation to justify actions to an authority (Cañas, 2022). Finally, there is *liability*, which denotes a form of legal responsibility defined by law, regulations, and/or contracts and therefore enforceable (Cañas, 2022). While acknowledging the importance of further examining liability issues in the context of human–AI collaboration, this dissertation deliberately sets aside the legal perspective, as it falls outside the scope of this work's specific focus on the self-attribution of responsibility for tasks and outcomes in human–AI collaboration by those individuals directly involved. More precisely, this dissertation focuses mainly on *moral responsibility* that is "backward-looking" and can be expressed through "blameworthiness" (Rowe et al., 2024, p. 592; Van De Poel, 2011).

Responsibility attribution describes the process by which someone evaluates "blameworthiness." In psychology, it is most commonly understood as a process through which individuals try to find causes for certain events (Cramton, 2002), which is influenced by people's perception of causality (Fincham & Jaspars, 1980). By building on prior works (including Heider, 1958; Rotter, 1966), Weiner (1985) introduced a psychological theory that explains how people interpret the causes of events, particularly their own and others' behavior in cases of successful and unsuccessful outcomes, and how these attributions determine future motivation, emotions, and behavior. While the theory has been applied mainly in human–human settings (e.g., Chandler et al., 1990; Greene et al., 2022), this dissertation adopts it for human–AI collaboration settings by arguing that AI systems are increasingly perceived as teammates (O'Neill et al., 2022), making people also apply "folk-psychological interpretation" toward humanoid technology (Thellman et al., 2017). Therefore, we argue that a deeper understanding of responsibility attribution processes in human–AI collaboration is essential to ensure such collaboration remains responsible, particularly given that AI systems themselves cannot assume responsibility (Swanepoel, 2021). While responsibility attributions can be directed toward others or toward oneself, I

refer to the latter in this dissertation as *self-attribution of responsibility* in human–AI collaborations—that is, "the extent to which [individuals] attribute responsibility [for outcomes] to themselves" in human–AI collaborations (Passlack et al., 2024, p. 5).

The **attributional theory** of Weiner (1985) explains how individuals infer causality for certain events, thus explaining responsibility attribution processes as presented in Figure 3. By building on prior work (Heider, 1958; Rotter, 1966), Weiner (1985) suggests that individuals experiencing certain outcomes (i in Figure 3)—either positive or negative—leads to an emotional reaction, such as sadness and disappointment or satisfaction. Following the emotional reaction, individuals engage in a cognitive process to identify possible causes for the outcome, which is shaped by causal antecedents (ii) such as biases or prior experiences. Correspondence bias and hedonic bias are particularly relevant to this dissertation. *Correspondence bias* refers to the tendency of individuals, when evaluating the behavior of others, to explain behavior by pointing to inherent traits or flaws (dispositional factors) rather than considering external circumstances (situational factors) that may have influenced the outcomes (Gilbert & Malone, 1995). In contrast, *hedonic bias* refers to the tendency of individuals to credit themselves (responsibility attribution to internal factors) for positive outcomes, while blaming outside circumstances (external factors) for negative ones (Weiner, 1979). Based on these influences, individuals form causal ascriptions for the outcomes (iii). The perceived causes for success and failure can be categorized along three key causal dimensions (iv): *locus*, that is, whether the cause is internal or external; *stability*, whether the cause is consistent over time or variable; and *controllability*, whether the cause is within the individual's control. Finally, these attributions have psychological effects (v), such as changes in self-esteem or motivation, and lead to behavioral consequences that influence future decisions and actions. For example, negative outcomes are often attributed to external factors such as task difficulty rather than personal failings, which protects self-esteem (Harvey et al., 1981; Weiner, 1985).

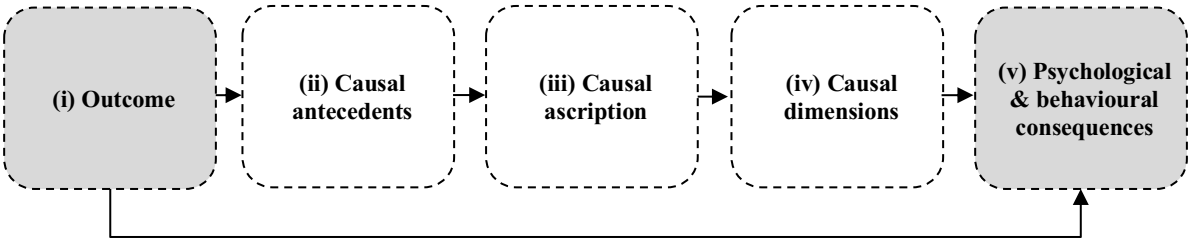


Figure 3: Key Elements of the Attributional Theory, Adapted from Weiner (1985) (Passlack et al., 2024)

Morality is a term relevant in the context of responsibility attribution, since responsibility often presupposes a moral framework. While in law responsibility is rather technical or formal, it has a moral dimension in this dissertation, as it is morality that guides human behavior and judgment (Barnes, 1999, p. 8). A *moral framework* can be understood as an understanding of what is wrong or right, defined through norms and often implicit rules (Barnes, 1999, p. 123). A *moral dilemma* arises when choices pit competing moral commitments against one another and each entails undesirable consequences (Lemmon, 1962). Assigning responsibility to someone, therefore, implies moral judgment about whether that person deserves praise or blame. Some argue that moral responsibility can be assigned only if individuals have control over their actions, act according to their own free will, and are able to understand the consequences of what they do (Coeckelbergh, 2020; Fischer & Ravizza, 1998). Having a high moral identity implies that morality—for instance, being honest and loyal—is part of a person's

character (Black & Reynolds, 2016). Scholars hold differing views on the nature of moral identity. Some, such as Frimer and Walker (2009), conceptualize it as a stable, trait-like characteristic central to an individual's self-conception. In contrast, others, such as Lapsley and Narvaez (2004), view it as a dynamic, context-dependent construct shaped by situational factors and deliberation. In other words, morality is not a single, uniform concept; rather, it is shaped by a variety of social and personal contexts, and thus is influenced by different ethical views (Vanhée & Borit, 2022) or "ethical families" (Bankins & Formosa, 2023).

Ethics is often described as the science of morality, as "the capacity to think critically about moral values and direct our actions in terms of such values" (Churchill, 1999, p. 259). Accordingly, different ethical frameworks can shape how moral values are analyzed and evaluated, leading to distinct judgments. While there are various ethical families, this dissertation differentiates between the most prominent ones, namely, normative ethics, meta-ethics, descriptive ethics, and applied ethics (Brendel et al., 2021). *Normative ethics* includes different streams such as deontology, consequentialism, virtue ethics, and contractarianism (Brendel et al., 2021; Kagan, 2018). While *consequentialism* judges the ethical rightness of an action by the results it produces (Frazer, 2006), *deontological ethics*, sometimes referred to as *duty ethics*, assesses whether actions are intrinsically right or wrong, independent of their outcomes; it thus assesses whether actions can become "universal law" as social norms or rules (Nowak, 2018, p. 189). *Contractarianism*, instead, holds that morality is grounded in mutual agreement within societies that seek to promote collective benefit. *Virtue ethics*, by contrast, focuses on the development of good character traits such as honesty, rather than rules (Bright et al., 2014; Flathmann et al., 2021); it is rooted in the philosophy of Aristotle. *Meta-ethics* is a stream that explores the nature, meaning, and foundations of moral concepts and judgments (Gewirth, 1960). *Applied ethics* or "*practical ethics*" instead aims to apply ethical theories and moral principles to particular real-world scenarios (Beauchamp, 2003). In contrast, *descriptive ethics* refers to empirical investigation of people's moral beliefs by observing, describing, and explaining how individuals make ethical considerations in real life. As Weaver and Trevino (1994) once summarized, while normative theories guide business ethicists in upholding moral values and principles, attention to descriptive ethics, grounded in empirical findings, helps researchers avoid proposing ethical initiatives that may appear sound in theory but prove impractical or counterproductive in practice. This underscores the dissertation's motivation to also learn from practice through empirical studies, following descriptive ethics.

Moral agency combines morality with ethics by defining the capacity of an entity to make moral judgments and be held responsible for its actions within an ethical framework (Mabaso, 2021). The debate over AI's moral agency is central to this dissertation. While some AI systems may demonstrate elements of ethical reasoning through emergent behaviors, they are perceived as lacking true autonomy or awareness, thus distinguishing their moral agency from that of humans (Coeckelbergh, 2020). Nevertheless, perceptions of AI having moral agency vary. AI may not be held morally accountable when it is viewed as a moral agent, depending on the intentions of its developers or users (Allen et al., 2000). However, when viewed as an autonomous agent capable of decision-making through self-learning, some may begin to attribute moral agency and responsibility to AI (e.g., Hong et al., 2020). And it can be argued that even without conventional moral agency, AI can affect human judgments and conduct (Susser, 2019). Differing interpretations influence how responsibility is assigned in practice and may lead to misconceptions about AI's capabilities to replicate human morality. As scholars have noted,

these uncertainties have wide-reaching implications for individuals and organizations (Y. K. Dwivedi et al., 2023; Pappas et al., 2023). In response, recent work has called for a deeper investigation into what shapes perceptions of responsibility, particularly among leaders engaging with algorithmic decision-making (McGuire & De Cremer, 2023).

Initial qualitative and pilot studies in human–AI settings suggest that individuals sometimes attribute responsibility to AI, particularly following unfavorable outcomes (e.g., Mayer et al., 2020; Passlack et al., 2023; Strich et al., 2021), and may blame AI systems even more than people in the context of autonomous driving (e.g., Hong et al., 2020). As AI systems are increasingly perceived as autonomous and human-like (Chandra et al., 2022), they are sometimes treated as teammates in collaborative settings (O'Neill et al., 2022; B. Zhang et al., 2021). At the same time, growing opacity of these systems (Berente et al., 2021) complicates assuming responsibility for users, and even developers often lack a full understanding of the processes, as models are often self-learning rather than explicitly programmed (Benbya et al., 2021; Berente et al., 2021). Although AI follows moral norms when they are encoded in its underlying model (Dyrkolbotn et al., 2018), we consider these entities to be accountable only if they act "by choice" (Barnes, 1999, p. 3; Coeckelbergh, 2020). Consequently, when AI systems generate decisions or suggestions, it can lead to responsibility gaps, particularly if individuals expect high degrees of accuracy (Hong et al., 2020), thereby potentially decreasing their critical reflection and encouraging misplaced attributions of responsibility to the AI system itself (Gunkel, 2020). As Gunkel framed it, "the question of social and moral status does not necessarily depend on what the other is in its essence but on ...how we decide, in 'the face of the other' ..., to respond" (Gunkel, 2018, p. 96).

Understanding how responsibility is attributed and perceived in organizational contexts where professionals collaborate directly with AI systems is expected to be critical for the long-term success of organizations. To illustrate this with an example, consider that when professionals attribute a successful outcome to their own decision-making, it likely fosters their self-efficacy and a sense of purpose. However, in the face of failure, responsibility may be attributed to the AI system as a means of protecting self-esteem (Harvey et al., 1981; Weiner, 1985). Findings from the literature reveal different views on whether AI can be perceived as having moral agency, which influences people's attribution of responsibility (Pansoni et al., 2023; Y. Zhang et al., 2023). This dissertation argues that if people acknowledge responsibility only in some cases, such as when decisions lead to success, that inconsistency coupled with moral uncertainty can lead to irresponsible behavior and, ultimately, organizational harm (Y. K. Dwivedi et al., 2023; Pappas et al., 2023). This understanding furthers the need for deeper investigations of perceptions of responsibility when collaborating with AI systems (McGuire & De Cremer, 2023). These considerations led to the fundamental question: ***How do professionals perceive responsibilities in human–AI collaborations? (RQ1)***

2.3 Changes in Ability Requirements

To address the second research objective (**RQ2**)—identifying the abilities and structural conditions that enable professionals to collaborate responsibly with AI—this section first delineates *AI literacy* as the conceptualization of knowledge, skills, and competencies needed to understand, use, and critically interrogate AI systems (Ng et al., 2021). Building on this foundation, I clarify adjacent notions of professionals' ability change: *reskilling*, understood as a structured, intentional shift in abilities to meet evolving demands in human–AI collaboration; and *upskilling* and *deskilling*, which capture more

informal changes in abilities (gains and losses) that arise in human–AI collaboration (Rafner et al., 2022).

AI literacy is a concept to which the literature most commonly refers when discussing required abilities for effective human–AI collaboration (Long & Magerko, 2020; Ng et al., 2021). *AI literacy* involves the ability to reflect critically on AI technologies, interact and work productively with them, and leverage AI as a practical resource in various contexts (Long & Magerko, 2020; Ng et al., 2021; Pinski & Benlian, 2023). As Ng et al. (2021) suggest, having AI literacy implies to "know and understand AI" which involve understanding the technical features of AI and how it works; "use and apply AI" by knowing how to and actually applying AI technologies in different contexts; "evaluate and create AI," which is linked to high-order thinking used to evaluate, appraise, and design AI; and employ "AI ethics," implying critical reflection upon AI's impact with respect to fairness, transparency, and so on. Many researchers argue for a strong connection between *digital literacy* and AI literacy (e.g., Anton et al., 2020; Wang et al., 2022), and some believe that being digitally literate is a necessary foundation for developing AI literacy (Anton et al., 2020; e.g., Jaiswal et al., 2022). This understanding needs to be extended by factors such as incorporating ethical considerations concerning AI's increased human-like abilities (Ng et al., 2021).

AI literacy "is a general concept adaptable to changing times and concerns" (Bawden, 2008, p. 23) and, like digital literacy, cannot be limited to a simple and static checklist of abilities, skills, and knowledge. It comprises a range of *skills* that are task-specific and can be acquired through training and experiential learning (Yamaguchi, 2018). It also encompasses *knowledge*, defined as the understanding of facts and information shaped by an individual's education and experiences (Dalkir, 2013, p. 8), meaning "information possessed in the mind" (Alavi & Leidner, 2001, p. 109). Knowledge tailored to a specific field or context is often referred to as domain-specific knowledge (Bradley et al., 2006); when it involves higher-order analytical awareness, such as knowing how or where to access relevant information, it is considered *meta-knowledge* (Evans & Foster, 2011). *Capabilities* emerge when individuals apply their knowledge and skills to respond effectively to external challenges (Wahl & Prause, 2013, p. 67). These capabilities are developed through experience and are also closely tied to a person's inherent talents and aptitudes (Ferguson, 1960, p. 131), which can be categorized as *personal dispositions*. In contrast, *competencies* are understood as the cross-functional integration and coordinated application of such capabilities across diverse contexts (Wahl & Prause, 2013, p. 67). This dissertation uses the term *abilities* as an umbrella term that includes the inherent or acquired capabilities of individuals to perform tasks, encompassing a combination of cognitive, emotional, social, and physical competencies that enable purposeful action and adaptation.

Reskilling is a concept linked to AI literacy, as it refers to acquiring new skills to meet the demands of a changing technological and societal environment—a dynamic process in which individuals expand their abilities into previously unfamiliar domains (Rafner et al., 2022; Sawant et al., 2022) to meet the changing requirements for collaborating with AI (Rafner et al., 2022). While AI literacy conceptualizes the skills relevant for collaborating with AI, and can thus inform training programs, practical experience shows that many competencies emerge through learning by doing, in the sense of non-formal learning (Ludwig, 2024, p. 27). Scholars distinguish formal, non-formal, and informal learning using three criteria: intentionality, the presence of explicit learning objectives, and institutional organization of the learning process (Werquin, 2010). *Formal learning* satisfies all three; *non-formal* typically involves

intentional learning without explicit objectives; and *informal lacks* all three, though views differ on whether informal learning can be intentional, and evidence suggests intentionality may be necessary for workplace impact (Jeong et al., 2018; Jurczyszyn, 2024). In this dissertation, informal learning denotes "learning by doing" that occurs without purposeful formal instruction or explicit intent to acquire new knowledge, but emerges incidentally while using AI to complete a task. Informal learning is a concept that may increase in importance given the fact that AI use in organizations is no longer introduced exclusively by management but often used informally and without supervision (Krakowski, 2025), such as to generate content (Y. K. Dwivedi et al., 2023), and thus triggering upskilling and deskilling processes.

Deskilling refers to the decline of certain abilities, such as critical reflection, as a result of automation or habitual reliance on AI (Chuang, 2020). In contrast, *upskilling* involves deepening existing skills to remain effective in familiar roles, such as learning to formulate precise prompts for generative AI tools (Moore et al., 2020; Rafner et al., 2022). It is evident that upskilling and deskilling cannot be viewed independently but are interrelated and mutually dependent concepts (Bravo, 2015). A decline in certain skills can lead to the development of new ones. This phenomenon is illustrated by the discussion of Orlikowski and Barley (2001), who suggest that users delegate their tasks to machines but simultaneously enhance their skills by becoming programmers of those machines. This transformation reveals that AI not only alters tasks but may also redefine professional roles and perceived responsibility, which, in turn, links back to RQ1.

While existing approaches that conceptualize AI literacy offer valuable insights into which skills and knowledge constitute AI literacy (e.g., Anton et al., 2020; Cetindamar et al., 2022; Heyder & Posegga, 2021; Ng et al., 2021; Pinski & Benlian, 2023; Wang et al., 2022) and thus promote effective human–AI collaboration, we see a need to fully account for (un)intended outcomes of human–AI collaborations by addressing responsibility considerations more than superficially (Passlack, Hammerschmidt, et al., 2025). While efficiency and utility maximization in organizations are important (Reinhardt, 1992), we argue that responsible human–AI collaboration aiming for employee wellbeing, fairness, and inclusion (Kegel & Ghanem, 2024) is important to guarantee long-term success (Benlemlih & Bitar, 2018). This highlights the importance of this dissertation, which takes responsibility as a central objective in order to better capture the ethical tensions and unintended consequences (Benbya et al., 2021). This fosters a more nuanced understanding of responsible human–AI collaboration, which has been called for in research (Trier et al., 2023).

A responsibility-oriented perspective foregrounds the abilities required for *responsible* human–AI collaboration and helps identify whether imbalances risk entrenching social or economic divides (Zamani & Vannini, 2025). While prior conceptualizations often emphasize the increased complexity of AI, which requires increasingly sophisticated technical skills, this dissertation shares the view that many technologies become more user-friendly as they become more broadly accessible to users, often by being open access, leading to requiring fewer specific technical skills (Maedche et al., 2019). Keeping this in mind, prior conceptualizations feature a strong focus on which technical knowledge is needed rather than adopting a more deliberate approach to what actually shapes human–AI collaborations to be responsible. This motivates this dissertation's other research question: ***What abilities and structural conditions enable professionals to engage in responsible human–AI collaboration?*** **[RQ2]**

3 Research Methods

This dissertation employs a multi-method approach to ensure a comprehensive investigation of the research objectives (**RQ1** and **RQ2**) by applying different methods (Reis et al., 2022) and combining insights from theory and practice, as emphasized by Burton-Jones et al. (2021). It is grounded in *systematic literature reviews* to establish a theoretical foundation and assess the current state of research (Section 3.1). In addition, it utilizes *semi-structured interviews and open-ended questions* for research phenomena that require an inductive, exploratory approach (Section 3.2). Furthermore, *experiments* are conducted, enabling the presentation of empirically supported relationships (Section 3.3).

Table 2 presents an overview of each paper's objectives, methods, and sample sizes. The subsequent sections build on Table 2 by elaborating on the methods employed in this dissertation.

Table 2. Overview of Methods and Related Papers

3.1 Systematic Literature Reviews				
Paper	**Objective**	**Method**	**Sample**	**Data Analysis**
Paper I	Understand the ethical management of human–AI collaborations	Descriptive review with theorizing as a primary goal (Paré et al., 2015)	42 articles	Deductive and inductive coding in MAXQDA (Mayring, 2014)
Paper II	Investigate how responsibilities in human–AI collaborations can be perceived	Descriptive review with synthesizing as a primary goal (Paré et al., 2015)	151 articles	Deductive and inductive coding in Microsoft Excel (Mayring, 2014)
Paper VI	Study what shapes AI literacy for responsible human–AI collaboration	Descriptive review with synthesizing as a primary goal (Paré et al., 2015)	46 articles	Deductive and inductive coding in MAXQDA (Mayring, 2014)
3.2 Semi-Structured Interviews & Open-Ended Questions				
Paper	**Objective**	**Method**	**Sample**	**Data Analysis**
Paper III	Explore how professionals attribute responsibility in human–AI collaborations	Open-ended questions collected through an online questionnaire	215 managers	Analysis conducted in MAXQDA; explorative in nature, drawing on the principles established by Gioia et al. (2013)
Paper V	Explore how perceived role responsibilities may change in human–AI collaborations	Semi-structured interviews, following the guidance of Myers and Newman (2007)	28 secondary teachers	Analysis conducted in MAXQDA; explorative in nature, drawing on the principles established by Gioia et al. (2013)
Paper VI	Study what shapes AI literacy for responsible human–AI collaboration	Semi-structured interviews, following the guidance of Myers and Newman (2007)	28 professionals	Analysis conducted in MAXQDA, following the deductive content analysis approach proposed by Mayring (2014)
Paper VII	Explore how human–AI collaborations transform abilities and ability requirements	Semi-structured interviews, following the guidance of Myers and Newman (2007)	12 education students	Analysis conducted in MAXQDA; explorative in nature, drawing on the principles established by Gioia et al. (2013)

Paper VIII	Explore why AI literacy may affect AI use in cases of AI errors	Open-ended questions collected through an online questionnaire	550 programmers	Analysis conducted in Microsoft Excel, following the deductive content analysis approach proposed by Mayring (2014)
Paper IX	Understand how AI literacy impacts perceptions of AI	Open-ended questions collected through an online questionnaire	151 professionals	Analysis conducted in Microsoft Excel, following the deductive content analysis approach proposed by Mayring (2014)
Paper X	Study how organizations can foster responsible human–AI collaborations	Semi-structured interviews, following the guidance of Myers and Newman (2007)	27 professionals	Analysis conducted in MAXQDA; explorative in nature, drawing on the principles established by Gioia et al. (2013)

3.3 Experimental Research

Paper	Objective	Method	Sample	Data Analysis
Paper III	Explore how professionals attribute responsibility in human–AI collaborations	Vignette experiments that outline empirically supported relationships (Atzmüller & Steiner, 2010)	215 managers	Statistical analyses conducted using the program R, applying a mixed-effects model to account for both fixed and random effects within the data Independent samples t-tests performed to compare mean differences across groups, allowing for initial evaluation of treatment effects Cronbach's Alpha (α) calculated to assess the reliability of measurement items for the dependent and control variables, following the guidelines of Petter et al. (2007) Reliability confirmed with α-values exceeding the 0.80 threshold, as recommended by Cho and Kim (2015; 2007) A two-way ANOVA employed to examine between-subjects effects related to the two treatment conditions, in accordance with Dennis and Valacich (2001)
Paper IV	Determine how professionals' responsibility attribution is influenced by their moral identity			
Paper VIII	Assess how AI literacy affects AI use in case of AI errors	Online experiment designed to examine empirically supported relationships	550 programmers	Linear mixed-effects models employed in the program R, following the framework proposed by Bates et al. (2024), to account for both fixed and random effects within the data

3.1 Systematic Literature Reviews

A literature review is "a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners" (Fink, 2019, p. 6). Literature reviews make "a vital contribution to the relevance and rigor of research" (Vom Brocke et al., 2009, p. 2). As a recognized scientific practice, they are considered a fundamental step in conducting research, providing the necessary foundation for any research study (Baker, 2000; Webster & Watson, 2002).

Systematic literature reviews are critical: they not only provide an overview of the existing knowledge about a topic under research, but also help in finding all relevant sources (Vom Brocke et al., 2009), outlining contradicting results and research gaps (Webster & Watson, 2002), and providing ideas and directions for future research (Rowe, 2014; Webster & Watson, 2002). While literature reviews can be seen as preparation for a research study, there is also the standalone literature review—"a scholarly paper that in its entirety summarizes and synthesizes knowledge from a prior body of research" (Okoli, 2015, p. 879). While literature reviews may serve various purposes—such as explanation, description, theory testing, or theory development to understand (Paré et al., 2015; Templier & Paré, 2018)—and

may lead to the development of ad hoc classification schemes, taxonomies, conceptual frameworks, or even theoretical systems (Webster & Watson, 2002), this dissertation employs them primarily to help describe and understand in preparation for subsequent empirical investigations.

Paper I is a theoretical review with theorizing as its primary goal (Paré et al., 2015) to explain how we can understand the ethical management of human–AI collaborations in organizations as sociotechnical systems through the lens of sociomateriality. This paper serves as the theoretical framework of this dissertation and its other papers. In contrast, **Paper VI** and **Paper II**, are descriptive reviews, as both aim to outline different perspectives, findings, conceptualizations, and trends of existing research and thus show the current state of research in the specified topics (King & He, 2005; Paré et al., 2015). The study in **Paper II** focuses on perceptions of moral agency and responsibilities in human–AI collaborations. **Paper VI** seeks to identify relevant abilities, competencies, and skills for collaborating with AI by synthesizing existing conceptualizations on AI literacy.

As Okoli (2015, p. 43) states, a "literature review must be *systematic* in following a methodological approach, *explicit* in explaining the procedures by which it was conducted, *comprehensive* in its scope of including all relevant material, and, hence, *reproducible* by others who would follow the same approach in reviewing the topic." The following section provides a detailed elaboration to address these key dimensions.

3.1.1 Sampling and Data Collection

The systematic literature reviews followed the steps suggested by Templier & Paré (2018). Prior to beginning the review process, the approach always had a conceptual stage to define the scope of the review and identify the constructs relevant to the study's aim. These depend on the different topics, which were then structured according to an integrating type of conceptualization (MacInnis, 2011) and resulted in an integrative framework. This means the keyword-based search followed a theoretical deductive approach by conducting preliminary searches and test queries to determine suitable databases and adjust the keywords that built the search string (Kitchenham & Charters, 2007). The keywords were chosen rather broadly as concepts, and all possible synonyms were included that were identified through engagement with the study context in order not to miss any relevant articles (Templier & Paré, 2018). After reviewing the initial articles, additional keywords were incorporated if new technical terms describing the same phenomena were identified. If no additional keywords were identified, the review stage began.

For the second step, the literature search, we searched for the identified keywords within the title, abstract, and keywords of each article. The literature was collected from IS-related databases, always including the most relevant and commonly used databases among IS scholars (Bandara et al., 2015): ACM Digital Library (ACM), AIS Electronic Library (AIS), Business Source Ultimate via EBSCOhost (EBSCO), IEEE Computer Society Digital Library (IEEE), JSTOR, ScienceDirect (SciDir), and Web of Science (WoS). In addition, the approach in **Paper I** was extended manually to include relevant journals not covered by these databases, such as the *Australian Journal of Labour Law*, *AoM Organizational Communication and Information Systems*, and ethics-related journals such as the *Journal for Business, Economics & Ethics* that are not covered in the IS rankings.

In the next step—screening for inclusion—we applied exclusion and inclusion criteria to all articles identified. These criteria can be split into format screening, quality screening, and relevance screening. The format screening guaranteed the exclusion of gray literature and incomplete articles, such as extended abstracts and research-in-progress, as well as standalone literature reviews lacking unique insights (e.g., those focused on scoping). The relevance screening excluded articles not relevant to the topic being researched by following predefined relevance criteria. The quality assessment was only applied in **Paper I** based on established rankings of the Harzing list. There was no application of quality screening for **Paper II** because there is no ranking in the discipline of ethics, and the aim was to guarantee that no relevant high-quality work was excluded; and for **Paper VI**, since the novelty of the construct "AI literacy" meant that applying quality screening would have excluded valuable conceptualizations that had thus far been presented only at conferences and were not included in the ranking.

The remaining articles deemed relevant were retained in the final sample and used to conduct forward- and backward-citation searches to minimize the risk of omission. After reaching apparent saturation, all identified literature was retrieved in full text and analyzed as detailed in the following section.

3.1.2 Data Analysis

The qualitative data analysis of prior literature always followed a deductive approach in the first step (Mayring, 2014), which was grounded in the concepts identified in the *conceptual stage* of the data collection (in Section 3.1.1). All relevant articles were obtained as PDFs for systematic reviewing. MAXQDA[4] was utilized for coding and analyzing in **Paper I** and **Paper VI**, whereas **Paper II** was examined using a structured coding approach in Microsoft Excel. The initial phase involved extracting meta-information from the articles, such as research domain, methodological approach, and key findings, which served as a foundational descriptive overview of the dataset.

Following Mayring's (2014) deductive methodological approach, the analysis began with the deductive development of a category system. The codebook included distinct categories, each defined by a name, a precise definition, coding rules, and anchor examples derived from the data. Approximately 30% of the material was initially coded in a structured "material run-through" phase. During this process, any new findings that did not align with existing categories prompted inductive category expansion. Subsequently, categories, definitions, coding rules, and examples were refined when uncertainties arose in the coding or when categories did not match the data. The material was recoded iteratively through multiple review cycles to ensure internal consistency. Ambiguities were addressed in that we, as coders, discussed and resolved discrepancies. Intercoder reliability was evaluated by parallel double-coding of about 20% of the data, thereby testing—and, where necessary, refining—the coding guideline. Following coding, the results were interpreted and analyzed. This analytic phase proceeded iteratively through collaborative discussions and structured brainstorming until consensus was achieved.

[4] www.maxqda.com

3.1.3 Data Quality, Validity, and Reliability

To ensure reproducibility, the literature selection process was conducted transparently and systematically, aiming for "literature review articles … to adhere to the same high standards of quality and trustworthiness as other empirical studies" (Paré et al., 2016, p. 493). Each paper details the search strategy, including the keywords used, as well as the inclusion and exclusion criteria applied. This allows other researchers to replicate the approach and obtain comparable literature findings (Kitchenham & Charters, 2007; Vom Brocke et al., 2009). To ensure data quality, we considered only peer-reviewed literature. In addition, publication outlets were evaluated based on their ranking within the Harzing list or by analyzing their h-index (in domains like ethics) to assess their credibility and impact. Adhering to a predefined codebook by following a deductive approach enhanced our rigor of the coding process (Mayring, 2014). The codebooks included clear definitions, illustrative examples, and specific coding rules "[w]here there are problems of delineation between categories" (Mayring, 2014, p. 95) to ensure consistency during coding and among coders. Cohen's Kappa (Cohen, 1960) was calculated to assess the reliability of the coding process by measuring the degree of agreement between coders in assigning text modules. In cases of disagreement, multiple discussion rounds were conducted among all coders to ensure clarity and consistency in the coding process.

3.2 Semi-Structured Interviews & Open-Ended Questions

Semi-structured interviews, which constitute a qualitative research approach, are often utilized to explore complex and insufficiently understood phenomena (Sarker et al., 2013). Qualitative research in general involves gathering in-depth data and learning from participants, often having an exploratory focus (Creswell & Creswell, 2018, p. 257; Marshall & Rossman, 2006). This method was applied when "the topic [was] new, the subject has never been addressed with a certain sample or group of people, and existing theories [did] not apply with the particular sample or group under study" (Creswell & Creswell, 2018, p. 57; Morse, 1991). It helped this dissertation to generate new theoretical insights and provide propositions for future research, and is today widely recognized as a valid approach within the IS research community (Sarker et al., 2013). In other words, "qualitative researchers can generalize from empirical observation by developing new concepts and propositions as a way toward theory building," (Conboy et al., 2012, p. 155) and are applicable to answer "open-ended," evolving, and "nondirectional" research questions such as "how" and "what" (Creswell & Creswell, 2018, p. 193). Qualitative research can also complement and contextualize the findings of quantitative studies (Creswell & Creswell, 2018, p. 41; Creswell & Poth, 2016, p. 40). While numerical data reveal patterns, trends, and correlations, they often lack insight into the underlying motivations, perceptions, and contextual factors driving human behavior (Creswell & Poth, 2016, p. 40). Qualitative research helped within this dissertation to mitigate this gap by examining individual perspectives (Sarker et al., 2013), thereby providing a deeper, more nuanced understanding of the "why" and enriching the interpretation of quantitative results.

In the case of **Papers V–VII**, and **Paper X**, the method is semi-structured interviews; while **Papers III, VIII,** and **Paper IX** draw on responses to open-ended questions collected through an online questionnaire. While it is sometimes argued that for qualitative studies "compared with quantitative studies, there is arguably less structural support" (Conboy, 2019, p. 115), this dissertation aimed to adopt well-established methodologies that emphasize systematic rigor and transparency as a way to navigate these challenges (Conboy et al., 2012). The methods applied are detailed in the following sections.

3.2.1 Sampling and Data Collection

In qualitative research, it is critical to deliberately choose participants based on their expertise and relevance to the research question (Palinkas et al., 2015) to obtain rich, insightful data. Whether an individual possesses expert knowledge was assessed jointly by the individual and the researcher(s). A key prerequisite for participation was direct involvement in the practical settings under investigation to ensure that participants had in-depth knowledge and nuanced insights into the processes being studied (Meuser & Nagel, 1989, p. 4). This approach aligns with purposeful sampling strategies, which involve "identifying and selecting individuals or groups of individuals that are especially knowledgeable about or experienced with a phenomenon of interest" (Palinkas et al., 2015, p. 534). Given the dissertation's focus on organizational contexts, the sample consisted primarily of different professionals, including managers, developers, and salespeople (**Paper VI,** and **Papers VIII–X**), while some had a specific focus on managers (**Paper III**, **Paper IV)**. **Paper VII** also included students who were about to enter professional roles, reflecting the dissertation's forward-looking perspective on emerging practitioners.

The interviews were semi-structured and followed the guidance of Myers and Newman (2007). As suggested in the literature, a semi-structured interview guide—sometimes referred to as "interview protocol (Creswell & Creswell, 2018, p. 266)—was used as "a schematic presentation of questions or topics" (Jamshed, 2014, p. 87). It guaranteed consistency and ensured comparability among interviews for this dissertation's findings, and ensured that no information was missed during the interviews (Creswell & Creswell, 2018, p. 266). Moreover, the interviews were useful for uncovering emerging phenomena (Creswell & Creswell, 2018, p. 41), and they offered the flexibility to adjust questions based on participants' answers to better capture the richness of the interviewees' experiences and perspectives, to draw conclusions from very subjective views. All interviews were recorded with the consent of the participants, which enabled us to focus on the flow of conversation without the distraction of note-taking, while also providing a reliable source for producing verbatim transcripts (Jamshed, 2014).

3.2.2 Data Analysis

Qualitative research allows a researcher to move from specific observations to broader themes while actively interpreting the data throughout the process (Creswell & Creswell, 2018, p. 41). The nature of the data analysis of this dissertation's qualitative papers was mainly inductive, although higher-order constructs and discussions were grounded in theory.

The data analysis was mainly conducted using the tool MAXQDA developed by Kuckartz (2014), which proved to be a suitable tool for qualitative analysis and organizing data (Mayring, 2014). Only the data in **Paper VIII** and **Paper IX** were analyzed using Microsoft Excel due to the scope of different participants; it was perceived as providing a better filter option for identifying patterns and to provide access to authors without a license for MAXQDA.

The analysis of the data in **Paper VI, Paper VIII,** and **Paper IX** was conducted following the content analysis approach proposed by Mayring (2014). Initially, a deductive procedure was applied, based on theoretically derived categories. Key theoretical components were defined to form the initial category system, followed by the selection of anchor examples from the data and the development of clear coding rules for each category. Where data could not be clearly assigned to these predefined categories, or

where content within a category proved highly heterogeneous, an inductive refinement of categories was undertaken (Mayring, 2014, p. 104).

The analyses in **Paper III, Paper V, Paper VII,** and **Paper X** were explorative in nature, meaning the process was exploratory and can be categorized as an interpretive form of qualitative research that mainly followed an inductive approach (Sarker et al., 2018). To ensure a rigorous and consistent analysis, a structured coding approach based on Gioia et al. (2013) was applied. In a first step, so-called "1st-order categories" were derived directly from the participants' own language, minimizing the influence of existing theory and preserving the authenticity of participant perspectives. To interpret these categories in light of broader conceptual patterns, the "2nd-order themes" were built, enabling a deeper explanation of the observed phenomena. Once the identified categories were grouped into themes, they were further abstracted into aggregate dimensions, grounded in theory for interpretation and formulating propositions.

3.2.3 Data Quality, Validity, and Reliability

A common critique of qualitative research is its limited generalizability, as the subjective nature of interviewed individuals' opinions can make it difficult to apply findings broadly (Gibbs, 2018; Sarker et al., 2013). In response to this and to ensure the research is trustworthy, we compared our findings with the literature to determine whether they were aligned or contradicted prior work. Doing so shows that results are credible and meaningful and grounded in a broader context (Gibbs, 2018). Also, to ensure no important data were lost, all interviews were fully transcribed from recordings rather than being based solely on researcher notes, which enhanced interpretive accuracy (Myers & Newman, 2007).

To enhance the trustworthiness of the findings, each step of the research process, from data collection to data analysis, is documented within the papers. This included a structured interview guide and coding guidelines to maintain consistent interpretation among all coders throughout the coding process and to strengthen the reliability of the analysis (Gibbs, 2018; Mayring, 2014; Sarker et al., 2013). To further strengthen the validity of the analysis, all coders engaged in joint coding of transcripts and developed and refined the coding scheme collaboratively through ongoing discussion (Garrison et al., 2006). Except for the single-author paper (**Paper V**), for each analysis approximately 20% of the data were coded by at least two coders. Intercoder reliability was assessed using Cohen's Kappa to control for interpretation bias (Cohen, 1960; McHugh, 2012), with the aim of achieving at least 80% agreement to ensure the reliability of the qualitative analysis (Miles & Huberman, 1994).

3.3 Experimental Research

Experiments are robust instruments for examining real-world scenarios while enabling the controlled manipulation of specific factors (Atzmüller & Steiner, 2010; Finch, 1987). Experiments fall under quantitative methods, using numerical representations of theoretical concepts (Orlikowski & Baroudi, 1991; Venkatesh et al., 2016) and analyzing data to understand the relationship between independent and dependent variables (Chen, & Hirschheim, 2004). Experiments for this dissertation ensured that individual behavior could be studied under consistent conditions, minimizing the influence of environmental variables like those in real-world settings, which could have obscured the findings (Steiner et al., 2016). Compared to traditional surveys, vignette experiments offered a significant

advantage by engaging participants in simulated situations where they actively reacted to a situation. This allowed for observing individual actual behavior rather than relying solely on self-reported intentions, as in traditional surveys (Atzmüller & Steiner, 2010).

This dissertation includes three papers that rely on experiments, **Paper III, Paper IV,** and **Paper VIII**. These papers tested hypotheses linked to professionals' advice-taking behavior and responsibility attribution when collaborating with AI in decision-making tasks. **Paper VIII** is an online experiment designed to examine how individual AI literacy affects utilization of AI in programming tasks, particularly in response to bad advice. In contrast, **Paper III** and **Paper IV** employ vignette experiments to explore responsibility attribution for decision-making outcomes in the case of failure and success and how this is influenced by professionals' moral identity. The various vignettes are short narratives that immerse participants in specific scenarios, requiring them to assume a role and respond to unfolding events—referred to as treatments (Finch, 1987). The dissertation thus provides insights into how individuals actually behave in human–AI collaboration settings, complementing the findings from what they recount regarding their perceptions (in interviews) or what theory posits they would do (from systematic literature reviews).

3.3.1 Sampling and Data Collection

Sampling and data collection took place on the click-worker platform Prolific in the case of all three papers, while the data collection for **Papers III** and **IV** was extended by a sample recruited via LinkedIn.[5] All experiments in this dissertation were conducted online, which has three major advantages: reaching a broader participant base; reducing costs; and collecting data despite location. Filter options in Prolific accounted for gender-balanced participation.

Concerning the selection of participant groups, the study of **Paper III** and **Paper IV** specifically targeted managers as key participants. High-quality decision-making is critical for organizational success and is profoundly influenced by managers' approaches to these processes (Dean & Sharfman, 1996). This is why it was imperative to engage practicing managers to ensure that respondents possessed direct experience with the complexities of real-world decision-making. To capture a holistic view of decision-making responsibilities, we recruited managers across different organizational levels and with diverse experience (Brendel et al., 2023). In **Paper VIII**, programmers deliberately constituted our experimental sample, as generative advances in AI have led to the growing adoption of generative AI to support programming (Stokel-Walker & Van Noorden, 2023) as a future skill. Using programmers as a sample allowed us to investigate the advice-taking behavior in programming when faced with imperfect advice.

The number of participants needed for each study was evaluated by conducting an *a priori* power analysis (Cohen, 1988; Faul et al., 2007) that allows for determining the required sample size N and offers a reliable approach to control statistical power prior to conducting experiments (Hager, 2006). The sample size N was determined based on the desired statistical power (1-β), predefined significance level (α), and expected effect size, which can be differentiated between small-, medium-, and large effect size (Cohen, 1988), ensuring a specified probability of detecting the effect (Cohen, 1988; Faul et al., 2007). Statistical power represents the probability of correctly rejecting the null hypothesis (H_0) when

[5] https://www.prolific.com; https://www.linkedin.com

it is indeed false. Tests with low statistical power have limited utility, as they fail to differentiate reliably between H_0 and the alternative hypothesis (H_1) (Faul et al., 2007). The power analysis resulted in N = 550 for the online experiment in **Paper VIII** and N = 215 for each vignette study on which **Paper III** and **Paper IV** rely.

The online experiment from **Paper VIII** was realized with oTree, while the experiment for **Paper III** and **Paper IV** was realized with SociSurvey.[6] All experiments began with the collection of sociodemographic data from the participants. For **Paper III** and **Paper IV** we did so to guarantee the suitability of participants, who needed to be managers or leaders, and to evaluate differences in outcomes depending on management level. In the case of **Paper VIII,** we conducted a brief questionnaire to ensure experience with programming, which was a prerequisite for participating in the experiment. Measurement items and open-ended questions were always collected after the experimental tasks to not influence participants by providing hints to the topic.

3.3.2 Data Analysis

Prior to data analysis, we applied specific exclusion criteria of participants in accordance with good practices (Aruguete et al., 2019). Participants were excluded from the dataset if they failed to answer all questions, did not pass attention checks, exhibited specific response patterns, completed the experiment in an implausibly short time frame, or exceeded the average completion time of other participants by more than two standard deviations.

The hypotheses in **Paper IV** were tested by conducting statistical analyses through linear regression models using the program R and applying a significance threshold of $p < 0.05$. More precisely, a two-way ANOVA was conducted to examine between-subjects effects in relation to the two treatment conditions (Dennis & Valacich, 2001). In **Paper III**, however, data analysis was conducted using mixed-effects models, appropriate for the repeated-measures design in which the dependent variable, responsibility attribution, was measured multiple times per participant. The data exhibited a nested structure, necessitating the use of this model to account for both within-subject factors, arising from repeated tasks, and between-subject factors, as participants were randomly assigned to distinct treatment groups (Atzmüller & Steiner, 2010). This approach allowed for group differences while controlling for individual-level variation. As a further control, t-tests were conducted to compare means among groups—more precisely, to compare the responsibility attribution among managers when facing positive outcomes versus when facing negative outcomes. The significance threshold was set at a p-value of ≤ 0.5. In the case of **Paper VIII** as well, the results rely on four linear mixed-effects models, which were conducted in the program R. Likewise, linear mixed-effects models (as proposed by Bates et al. (2024)) were chosen, as they enable varying intercepts for the different assignments (meaning that each assignment can have its own baseline level) and account for the assignment design. For all papers, the reliability of the measurement items for the dependent and control variables was tested using Cronbach's Alpha α (Petter et al., 2007) to guarantee the results demonstrate reliability by surpassing the 0.80 level (Cho & Kim, 2015).

[6] www.otree.org; www.soscisurvey.de

3.3.3 Data Quality, Validity, and Reliability

To ensure reliability of results, hypotheses for the experiments were grounded in established theoretical frameworks and were preregistered, in the case of **Paper III** and **Paper IV**, via OSF Registration and in the case of **Paper VIII** via registered Report at BISE.[7] This helped enhance transparency and prevent *post hoc* modifications based on the data outcomes. The experimental design used in **Paper III** and **Paper IV** received approval from the Ethics Committee of the University of Bamberg, ensuring compliance with ethical standards. Concerning participant selection and to account for external validity in **Paper III** and **Paper IV**, two different online channels were used to recruit participants (Prolific and LinkedIn), and participants were randomly assigned to the two research groups. *A priori* power analysis guaranteed the findings have a certain power, and aptitude checks were used to evaluate and monitor the quality of answers (Aguinis et al., 2021; Coppock & McClellan, 2019). To guarantee internal validity, easily understandable language was used for the tasks, and there was always a pilot study to test the feasibility of the design (Aguinis et al., 2021; Boudreau et al., 2001). Moreover, multiple robustness checks were conducted to evaluate and ensure the validity and reliability of the findings. The average time of the experiments was tested through the pilot study, and participant compensation was set slightly above the suggestion made by Prolific (which represented the higher end of minimum wages in the US states). Open-ended questions were always placed at the end of the experiment to avoid participants being more sensitized to the topics under research before completing the tasks of the experiment. In the case of **Paper III** and **Paper VIII,** open-ended questions were asked to complement quantitative analysis in the sense of a mixed-methods design, which involves collecting qualitative and quantitative data simultaneously and merging them to provide comprehensive analyses, with any contradictions explored or explained during interpretation and thus increasing internal validity (Reis et al., 2022).

[7] https://osf.io/; https://bise-journal.com/?page_id=2258

4 Key Findings and Contributions

In response to the dissertations' two research objectives, the following sections build upon the theoretical framework based on **Paper I** (section 2.1) as a theoretical foundation and present the key findings and contributions from the other nine papers. While **Papers II** through **V** offer a deeper exploration of perceived responsibilities in human–AI collaborations, thereby contributing to RQ1 (see Section 4.1), **Papers VI** through **X** address RQ2 by investigating the abilities and structural conditions that enable professionals to engage in responsible human–AI collaboration (see Section 4.2).

4.1 How do Professionals Perceive Responsibilities in Human–AI Collaboration? [RQ1]

Four papers contribute to RQ1 regarding perceptions of responsibility in human–AI collaborations. Building on a theoretical investigation of how individuals may perceive the responsibilities of AI-based technologies (**Paper II**), the dissertation offers empirical insights into responsibility attribution in AI-advised decision-making scenarios (**Paper III**) and examines how these perceptions are influenced by individuals' morality (**Paper IV**). It further provides initial findings on how collaboration with AI-based technologies may shape perceptions of professional role responsibilities in the educational context (**Paper V**). Table 3 is an overview of each paper's key findings that address the first research objective, and the subsequent pages offer a detailed description of these findings.

Table 3. Key Findings and Contributions Related to RQ1

How Do Individuals' Perceive Responsibilities in Human–AI Collaboration?			
Paper	**Research objective**	**Method**	**Key Findings & Contributions**
Paper II	Investigate how responsibilities in human–AI collaborations can be perceived	Descriptive review, with synthesizing as a primary goal (Paré et al., 2015)	• Attribution of moral responsibility toward AI is highly dependent on different ethical views • Distinctions between explicit and implicit ethical agents highlight differing views of moral reasoning in AI • Legal responsibility continues to rest solely with humans • Recent research advocates examining the full lifecycle of AI deployment, encouraging distributed responsibility among stakeholders
Paper III	Explore how professionals attribute responsibility in human–AI collaborations	Vignette experiments that outline empirically supported relationships (Atzmüller & Steiner, 2010)	• Individuals feel less responsible for negative outcomes (compared to successful ones) when making decisions collaboratively with AI, revealing a hedonic bias and a persistent responsibility gap • In cases of successful outcomes, individuals tend to feel more responsible when the advice comes from AI rather than from another person • Individuals perceive AI as requiring more justification for its advice to be seen as legitimate
Paper IV	Examine how professionals' responsibility attribution is influenced by their moral identity	Vignette experiments that outline empirically supported relationships (Atzmüller & Steiner, 2010)	• Individuals' moral identity has a positive effect on self-attribution of responsibility in decision-making conducted collaboratively with AI, regardless of outcomes • Moral identity does not significantly reduce the hedonic bias, implying greater attribution of responsibility to AI when human-AI collaboration has negative outcomes
Paper V	Explore how perceived role responsibilities may change in human–AI collaborations	Semi-structured interviews, following the guidance of Myers and Newman (2007)	• Generative AI is perceived and used in different ways: as a supportive aid; critical counterpart; disruptive influence; and sparring partner. • Generative AI use may lead to different professional role responsibilities: cheerleading; consulting; enlightening; and challenging • The interviews suggest four propositions regarding how AI perceptions correspond to particular professional role responsibilities

4.1.1 How Responsibilities in Human–AI Collaborations Can be Perceived [Paper II]

Paper II lays the groundwork for understanding how human–AI collaboration may influence individuals' perceived responsibility, as it summarizes existing literature on different perceptions of whether AI has moral agency and can be ascribed moral responsibility. It responds to the frequently discussed blurred lines of accountability that exist mainly due to the opacity of algorithms (e.g., Giermindl et al., 2022; Martinho et al., 2021) and their increasing autonomy and learning abilities (Baird & Maruping, 2021; Berente et al., 2021). **Paper II** contends that technological advancements necessitate a rethinking of responsibility in task-solving contexts, given the presence of autonomous AI-based entities that cannot bear responsibility in a legal sense (Swanepoel, 2021). Acknowledging that morality is not monolithic, but rather shaped by people's diverse ethical perspectives, the paper examines how moral agency and responsibility can be conceptualized and understood in human–AI collaborations.

Analyzing 151 scholarly articles resulting from a systematic literature review, **Paper II** found different perceptions regarding the moral agency of AI. Some described AI technologies as explicit ethical agents (Moor, 2006) that operate strictly within predefined ethical parameters without autonomous moral reasoning, as they lack consciousness (Roberts & Montoya, 2022) and emotional capacity (e.g., Giroux et al., 2022). This aligns with scholarly views, linked to deontological ethics, that AI systems merely execute programmed principles rather than exercise genuine moral agency (e.g., Kim et al., 2018; Taeihagh, 2021). Others, conversely, perceive AI technologies as implicit ethical agents (Moor, 2006) that develop ethical reasoning through unsupervised learning from data inputs, even in the absence of explicitly coded moral rules—a view linked to virtue ethics. This emergent functionality complicates human oversight, raising debates over whether such systems warrant classification as "moral agents" (Allen et al., 2000, p. 354). Hauer (2022, p. 1) articulates a central argument in this case, positing that the full automation of tasks requiring human moral authority inherently necessitates delegation of moral competence to autonomous systems. This underscores the ethical implications of AI conducting tasks autonomously, ascribing to AI responsibilities traditionally attributed to humans.

Moreover, **Paper II** outlines that to ensure that AI acts responsibly, ethics should be programmed by design (Horváth, 2022), which, moreover, can help people adhere to ethical guidelines (Schlimbach & Khosrawi-Rad, 2022) because AI will be sticking to predefined rules. Implementing ethics in AI design places responsibility on developers for AI's behavior (J. Zhang & Zhang, 2023). No single stakeholder group, though, can bear sole responsibility for the outcomes of human–AI collaboration, though, as many agree (Khan et al., 2022), since human–AI collaboration is a process in which different stakeholders—developers, managers, and users—are involved at different stages (Sullivan & Fosso Wamba, 2022). Palmer and Schwan (2024) argue for focusing more on the processes involved in AI development and use, rather than just debating abstract principles such as transparency and fairness. They emphasize the importance of understanding how people perceive morality throughout the entire AI lifecycle and highlight the need for further insights into virtue ethics.

Given the literature's predominant emphasis on deontological ethics, **Paper II** concludes that achieving a cohesive understanding of responsibility perception in human–AI collaborations will require a dialogue across disciplines and discussions among individuals with diverse ethical perspectives. To avoid responsibility gaps arising from increasing AI autonomy, the paper emphasizes the need for clarity in delineating the extent of human involvement and establishing AI accountability mechanisms that clarify the responsibilities of developers, users, and other stakeholders. The paper agrees with prior

research, such as that of Palmer and Schwan (2024), that we need empirical studies to look more closely at how people think and behave when collaborating with AI, moving beyond purely philosophical debates on morality or a sole focus on legal considerations. This further emphasizes the need for the investigations of **Paper III** and **Paper IV** described in the following.

4.1.2 How Professionals Attribute Responsibility in Human–AI Collaborations [Paper III]

Paper III contends that organizations are turning more frequently to algorithmic recommendations in decision-making, attracted by their perceived objectivity and their capacity to analyze vast datasets (Gunaratne et al., 2018) without emotional bias (Giroux et al., 2022). Research highlights the importance of human judgment, however, especially in decisions impacting employees' lives (Gerdes, 2022), and raises concerns about how algorithmic input affects individuals' sense of responsibility when outcomes fall short (Mayer et al., 2020; Passlack et al., 2023). As AI systems gain in autonomy and opacity (Angerschmid et al., 2022), especially through machine learning and large language models, they increasingly act as perceived "teammates" in decision-making, which can challenge human understanding and control (Angerschmid et al., 2022; Hakli & Mäkelä, 2019). This shift blurs the lines of responsibility (Giermindl et al., 2022), as professionals may feel compelled to follow AI advice without fully understanding it. This intransparency may lead to ambiguity with respect to who to blame (AI or people) and potential liability issues (Gunkel, 2020), which becomes particularly relevant in cases of unintended consequences such as errors or discrimination. Scholars argue that to ensure responsible AI use, transparency measures such as explainable AI must be adopted (Iyer et al., 2018), but these are only effective if users are willing to assume responsibility in human–AI collaborations.

Based on vignette experiments (with 108 managers solving four managerial decision-making tasks, resulting in a total of 432 observations) advanced by open-ended questions, **Paper III** finds that while managers perceive that AI cannot be responsible for decision-making outcomes, they feel significantly less responsible for unsuccessful outcomes compared to successful ones when making decisions collaboratively with AI. This highlights a responsibility gap, as critical issues related to accountability remain unresolved when managers perceive that they no longer hold responsibility for their actions. The paper further demonstrates that there is no significant difference in the effect described between managers who received advice from humans compared to those who received advice from AI. This suggests that managers tend to treat technological input similarly to human input, despite the fact that technology is not perceived as bearing responsibility. It is interesting that a significant difference emerged in cases where the outcomes were positive: managers who received advice from AI reported feeling more responsible than those who received advice from a person. The qualitative analysis aimed to complement the existing findings by exploring potential reasons for differences in self-attribution of responsibility when managers receive advice from AI versus human advisors. The insights suggest that AI agents require more extensive explanations than human advisors to convince managers of the legitimacy of their advice. Also, there is a preference for human advisors when decision outcomes directly affect individuals. Moreover, while managers acknowledge that AI lacks moral agency, they also recognize its potential to support moral considerations within collaborative decision-making processes. **Paper III** thus concludes that there remain responsibility gaps, as managers tend to attribute less responsibility to themselves for unfavorable outcomes in collaborative decision-making despite acknowledging that AI lacks full agency, consciousness (Roberts & Montoya, 2022), and free will

(Barnes, 1999). Moreover, the paper concludes that differences in how responsibility is attributed may be influenced by individuals' characteristics, such as their level of experience or the decision-makers' moral intention.

4.1.3 How Professionals' Responsibility Attribution is Influenced by Their Moral Identity [Paper IV]

We investigated in **Paper IV** the extent to which the responsibility gaps found in **Paper III** in cases of negative outcomes from human–AI collaboration can be mitigated by individuals' sense of morality. Prior research refers to moral identity to mean how strongly people see being moral as part of who they are (Black & Reynolds, 2016; Hardy & Carlo, 2011), based on their inherent beliefs of what seems right (Reynolds & Ceranic, 2007). A strong moral identity indicates that individuals are likely to behave honestly and avoid dishonest behaviors such as cheating, reflect on their actions, and admit their mistakes (Black & Reynolds, 2016). Thus, moral identity is suggested to play a key role in shaping how professionals make decisions and how responsible they feel for the outcomes (Heyder et al., 2023).

Accordingly, **Paper IV** investigated, based on vignette experiments, the extent to which leaders' moral identity influences their responsibility attribution toward themselves when their reliance on algorithmic advice leads either to successful or unsuccessful outcomes. The paper presents empirical evidence indicating a positive relationship between moral identity and self-attribution of responsibility. While it supports the existence of a hedonic bias, implying individuals attribute less responsibility to themselves for negative outcomes than positive ones, **Paper IV** does not find evidence that moral identity significantly moderates this effect by reducing the impact of negative outcomes on self-attribution of responsibility. One possible explanation lies in the relatively uniform composition of our sample, which comprised only managers exhibiting consistently high levels of self-reported moral identity, potentially constraining the variability necessary to observe significant moderating effects. It is also conceivable that participants' decision-making tasks were perceived as lacking real moral relevance given the experimental setting (De Freitas & Johnson, 2018). Yet, the paper concludes that moral identity is a critical factor to consider in fostering responsible human–AI collaborations, as it has direct effects on individuals' willingness to assume responsibility for decision outcomes.

4.1.4 How Perceived Role Responsibilities May Change in Human–AI Collaborations [Paper V]

Paper V advances the discussion on perceived responsibility in human–AI collaborations by focusing on one group of professionals, secondary school teachers, who play a crucial role in shaping the professionals of tomorrow. It investigates how embedding generative AI in educational settings reshapes teachers' perceptions of their role responsibilities and professional identities. The paper argues that generative AI— computer systems that learn from existing data and can create "new, meaningful content such as text, images, or audio" (Feuerriegel et al., 2024, p. 111)—is becoming increasingly democratized in the workplace (Krakowski, 2025) and has the potential to significantly alter employee roles and systems of responsibility (Giermindl et al., 2022; O'Neill et al., 2022). Such changes, in turn, may profoundly affect professionals' sense of meaningfulness at work (Harder Fischer et al., 2024) and impact perceived role responsibilities.

Rather than speculating from an external perspective, **Paper V** foregrounds the voices of teachers already engaging with generative AI in their everyday practices. Drawing on 28 semi-structured

interviews with secondary teachers, it examines not only how they reconfigure their teaching practices in response to generative AI but also how they perceive and interpret shifts in their professional role responsibilities. The analysis identified four distinct perceptions of generative AI, each linked to a corresponding role responsibility, resulting in four propositions. Teachers who perceive generative AI primarily as a supportive tool tend to emphasize their responsibility for motivating students. Those who experience generative AI as a disruptive presence tend to view their responsibility as enlightening students about AI's implications for them and society. Teachers who see generative AI as a critical friend tend to frame their responsibility as advising students on how to use generative AI effectively in their learning processes. Finally, those who engage with generative AI as a sparring partner perceive their role responsibility as challenging the reliability of AI and critically scrutinizing its outputs. **Paper V** contributes to IS research by specifically highlighting how teachers' professional role responsibilities may evolve and thus responding to recent calls for more research on potential digital futures (Schlagwein et al., 2025).

4.2 What Abilities and Structural Conditions Enable Professionals to Engage in Responsible Human–AI Collaboration? [RQ2]

Five of this dissertation's papers aim to answer RQ2 regarding the abilities and structural conditions that enable professionals to engage in responsible human–AI collaboration, with a focus on the individual and organizational levels. These papers suggest how we need to extend the concept of AI literacy when aiming for responsible human–AI collaboration (**Paper VI)**; explore how abilities may develop through collaborating with AI (**Paper VII)**; and provide first insights into how AI literacy may actually affect individuals' AI utilization (**Paper VIII**) and perceptions of AI (**Paper IX**). Finally, **Paper X** makes the dissertation's first suggestions for how organizations may foster responsible human–AI organizations among professionals by responding to ethical dilemmas.

Table 4 summarizes the key findings of each paper contributing to the second research objective of the dissertation, and the following pages elaborate on their findings.

Table 4. Key Findings and Contributions Related to RQ2

What Abilities and Structural Conditions Enable Professionals to Engage in Responsible Human–AI Collaboration?			
Paper	**Research Objective**	**Method**	**Key Findings & Contributions**
Paper VI	Study what shapes AI literacy for responsible human–AI collaboration	Descriptive review, with synthesizing as a primary goal (Paré et al., 2015)	• Conceptualization of AI literacy for responsible human–AI collaboration requires deeper considerations of personal dispositions, a focus on human-centered aspects and ethical literacy, and consideration of influencing factors • Problems concerning skilling and responsibility may challenge responsible human–AI collaboration • Propositions are to conceptualize AI literacy as a toolbox, consider the concept of multi-literacy, and reflect on shared responsibilities
Paper VII	Explore how human–AI collaborations transform abilities and ability requirements	Semi-structured interviews, following the guidance of Myers and Newman (2007)	• Individuals perceive all three—upskilling, deskilling, and reskilling—in human–AI collaborations • Critical reflection in human–AI collaborations is a skill gap • There are divergent findings concerning creativity: AI use is seen as both enhancing and diminishing creativity • Motivation may influence perceptions of ability development, with intrinsically motivated individuals more likely to fear deskilling than those driven by external incentives

Paper VIII	Consider how AI literacy affects AI use in case of AI errors	Online experiment designed to examine empirically supported relationships	• Individuals with low AI literacy are more willing to follow generative AI advice in programming contexts, despite the quality of the advice • Higher levels of AI literacy promote greater reflection on the potential limitations of generative AI • Exposure to bad AI advice exerts lasting effects on advice-taking, reducing uptake even when subsequent advice is accurate; this effect is not moderated by AI literacy
Paper IX	Develop insights into how AI literacy impacts perceptions of AI	Open-ended questions in online survey	• Varying levels of AI literacy shape individuals' perceptions of AI assistants: • Individuals with low AI literacy tend to perceive fewer distinctions between AI and human intelligence • AI literacy is needed to shape informed and more realistic perceptions about AI assistants
Paper X	Suggest how organizations can foster responsible human–AI collaboration	Semi-structured interviews, following the guidance of Myers and Newman (2007)	• Four recommendations on how organizations may respond to ethical decision dilemmas: o Strengthen professionals' reflection abilities to integrate diverse perspectives and balance shared responsibility with clear individual accountability o Foster moral values through cross-functional collaboration among lawyers, managers, and technicians, supported by structured dialogue to align differing ethical viewpoints o Embrace diversity in processes, particularly in content-generation tasks, as a way to navigate the tension between standardization and individualized approaches o Provide training on AI's limits and communicate transparently about its appropriate uses; support trial-and-error learning to build trust, while preventing overtrust through empathetic leadership and compliance oversight

4.2.1 What Shapes AI Literacy for Responsible Human–AI Collaboration [Paper VI]

Paper VI seeks answers to what shapes responsible collaboration between professionals and AI by investigating the required abilities and factors that influence their development. The paper acknowledges that prior research is split into research focused on the skills and knowledge needed to enhance technological efficiency and research that explores how AI affects humans as social entities (Bankins et al., 2024). By applying a lens of digital responsibility, accounting for the reflection of unintended consequences of human–AI collaboration (Trier et al., 2023), the paper seeks to identify the abilities and factors that shape responsible, or in contrast, irresponsible human–AI collaboration that goes beyond efficient outcomes, looking toward responsible outcomes. It argues that the increased technological usability that enables widespread use regardless of advanced technical skills (Maedche et al., 2019) does not necessarily ensure responsible collaboration. Therefore, it aims to extend the concept of AI literacy to guarantee responsible human–AI collaboration.

Building on a systematic literature review that synthesizes existing conceptualizations of AI literacy, **Paper VI** proposes a refined conceptualization and then extends it through a responsibility-oriented lens. The analysis draws on 28 interviews with people in five professional groups (HR professionals, developers, educators, salespeople, and managers) and provides four key findings that nevertheless acknowledge that some tensions reflect differing views among professionals. First, the findings indicate that personal dispositions become increasingly relevant when aiming for responsible human–AI collaboration. While some suggest that personal dispositions are not learnable, others expect them to be shaped and refined through ongoing human–AI collaboration, meaning they can develop. Second, it suggests that as AI systems grow in complexity, they demand greater technical expertise; conversely, some suggest that enhanced usability reduces the need for technical skills but elevates the importance of domain-specific knowledge and communication skills. Third, the paper outlines that ethical literacy can be viewed as a foundational element of AI literacy, yet AI literacy may also serve as a prerequisite

for developing ethical literacy, suggesting a bidirectional relationship. Fourth, it finds that deskilling is driven primarily by micro-level factors, whereas upskilling is more frequently influenced by meso- and macro-level factors.

To address the key tensions, the paper offers four actionable recommendations for practice. First, it suggests cultivating a shared sense of responsibility among professionals by facilitating discussions on appropriate and inappropriate AI uses across different contexts, emphasizing ethical reasoning. Second, it recommends conceptualizing AI literacy as a dynamic "toolbox," adaptable to diverse human–AI interaction scenarios while reinforcing shared accountability. The toolbox concept suggests that every use case demands a tailored mix of specific elements and foundational components, shaped by moderating literacies and factors from different levels of analysis. Certain literacy components are interdependent; for example, responsible human–AI collaboration requires the combined application of ethical considerations and development principles, underscoring the interconnected nature of these elements. Shared responsibility means that developers are accountable for embedding ethical considerations when making algorithmic design choices, while users need to reflect critically on outputs and ensure the system is used responsibly. Third, **Paper VI** recommends shifting the focus of an organization's training from solely technical skills to nurturing moral values and a commitment to lifelong learning. Fourth, it suggests targeted training programs by identifying the most critical competencies for specific professional roles and encouraging peer-based learning.

Linked to the four key findings, the paper also presents four propositions for future research. The first is that professionals' personal dispositions shape their sense of responsibility for AI-advised outcomes, with deficits in certain dispositions potentially leading to responsibility gaps. Second, the interplay between technical understanding and domain expertise affects responsible human–AI collaboration, with the relative importance of each varying by AI system and task context, posing challenges for aligning responsibilities with professionals' abilities. Third, the synergistic cultivation of technological and non-technological literacies has a greater impact on fostering responsible human–AI interaction than advancing any single literacy in isolation. Fourth, increased AI user-friendliness can lead to deskilling as individuals rely more on AI and less on their own competencies, thus eroding the skills necessary for responsible collaboration.

4.2.2 How Human–AI Collaborations Transform Abilities and Ability Requirements [Paper VII]

The objective of **Paper VII** was to investigate, from the perspective of informal learning processes, individuals' developments in skilling that come with increased collaboration with AI. While prior research mainly focuses on which skills are required for collaborating with AI (as did **Paper VI**), thereby implying approaches to formal learning, this study was interested more in perceptions of ability development triggered through informal learning processes that may foster or decrease certain abilities, linked to the concepts of deskilling, upskilling, and reskilling as introduced in the theoretical background. Recent research suggests that 70% of learning happens through non-formal learning (Ludwig, 2024). However, existing research focuses mainly on formal learning processes, despite the increase in open-source applications used by professionals that are not introduced by management (Y. K. Dwivedi et al., 2023; Krakowski, 2025). Thus, **Paper VII** seeks to contribute to research by providing insight into changes in abilities through the informal learning process, with the goal of informing studies

on formal learning strategies with respect to how abilities may develop in human–AI collaboration. This can subsequently guide the development of curricula for formal education and training programs.

Based on semi-structured interviews with students who employed generative AI in their research projects, **Paper VII** explored two dimensions of skill development in human–AI collaboration: the enhancement of existing abilities (upskilling) which were communications skills, digital competencies and broader domain knowledge, and the decline of certain skills due to reliance on AI (deskilling), such as creative thinking, critical reflection, and learning abilities. Also, it asks students about reskilling requirements. A key finding highlights a significant skill gap—while individuals acknowledge the importance of critical reflection, they often report a decline in this ability when depending on generative AI tools. It is interesting to note the different perceptions on AI's impact on creativity: some view AI use as expanding their idea generation and variation through brainstorming, while others contend it decreases originality by encouraging dependence on system outputs. Furthermore, the study reveals that motivational factors may influence individuals' perceptions of their skill development: intrinsically motivated students tend to express greater concern about deskilling, whereas extrinsically motivated students are more likely to perceive upskilling. The paper concludes that fostering responsible human–AI collaboration requires a deeper understanding of skill development dynamics to prevent an "illusion of knowledge" and ensure the responsible use of AI.

4.2.3 How AI Literacy Affects AI Use in Case of AI Errors [Paper VIII]

Paper VIII argues that while the literature stresses the importance of AI literacy for effectively collaborating with AI (e.g., Cetindamar et al., 2022; Pinski et al., 2023), research is mixed on the extent to which AI literacy leads to better outcomes in human–AI collaborations. While some research links AI knowledge (as a facet of AI literacy) to greater reliance on AI (through higher delegation of the task to AI (Pinski et al., 2023) and more advice-taking (Panigutti et al., 2022)), much of this evidence comes from studies in which the AI's suggestions were disproportionately good (e.g., Tully et al., 2025). In contrast, people who possess strong statistical literacy were found to tend to trust algorithms less (Marmolejo-Ramos et al., 2025). Limited AI understanding can inflate perceived AI intelligence (Ehsan et al., 2021), thus increasing advice-taking. This debate gains significance in light of the fact that generative AI systems are inherently imperfect (Lebovitz et al., 2021) and can generate inaccurate or misleading outputs commonly referred to as "hallucinations" (Maynez et al., 2020) due to their reliance on existing data.

Thus, **Paper VIII** investigates how AI literacy influences individuals' collaboration with AI**,** specifically, their willingness to accept advice from generative AI in programming tasks when that advice is incorrect**,** and how this impacts their future advice-taking behavior. The study employs an experimental design complemented by open-ended questions to gain deeper insights into the underlying reasons for participants' advice-taking. These open-ended responses are used to examine the extent to which AI literacy helps mitigate correspondence bias when individuals seek explanations for having received bad advice. The quantitative analysis shows that, overall, individuals with lower levels of AI literacy are more likely to follow advice from generative AI. Also, the data did not support the assumption that AI literacy lessens the impact of bad advice on future advice-taking: for individuals with both high and low levels of AI literacy, advice-taking decreased after AI had provided bad advice. In fact, individuals with higher AI literacy were generally more cautious and took less advice from the

generative AI, regardless of whether the advice was good or bad. The qualitative analysis added to this by revealing that AI literacy supports clearer and more confident reasoning about AI's performance. Moreover, the findings indicate that higher levels of AI literacy promote skepticism toward generative AI and greater reflection on its potential limitations, while individuals with lower AI literacy tend to hold unrealistically high expectations, suggesting the presence of correspondence bias.

4.2.4 How AI Literacy Impacts Perceptions of AI [Paper IX]

Paper IX extends the preliminary findings from **Paper VIII**, suggesting that AI literacy affects perceptions of AI in general. It presents findings on how individuals' AI literacy actually influences their perceptions of AI, which seems relevant given that perceptions influence behavior (Weiner, 1979). The paper argues that responsible human–AI collaboration depends on both the necessary abilities and knowledge about AI (referred to as AI literacy), as well as realistic perceptions of AI to balance trust and overtrust. This combination ensures that people can properly understand the limitations of AI and its potential impact on individuals and society at large. To address inflated views of what AI can do, the paper seeks to explore how professionals with different levels of AI literacy perceive AI. In doing so, the paper addresses the call to take a closer look at the differences between overly optimistic (utopian) and overly negative (dystopian) narratives about AI (Carroll et al., 2024; Koukouvinou & Holmström, 2022) and proposes how individual AI literacy may impact those perceptions.

Paper IX relies on data from an online survey combining closed and open-ended questions, with 151 valid responses. The closed questions served to measure individuals' AI literacy; the open-ended questions were used to analyze participants' narratives about AI's potential, functioning, and impacts, as narratives serve as communicative expressions of how they perceive AI (Cave & Dihal, 2019). The qualitative analysis revealed differences in how individuals perceive AI depending on their AI literacy level. Participants with stronger AI literacy showed more nuanced and objective understandings of AI, recognizing both its advantages and limitations. In contrast, those with lower AI literacy tended to frame AI in more negative narratives when reflecting on its implications. While AI-literate individuals emphasized the need for further training to engage responsibly with AI systems, participants with lower literacy levels often expressed expectations for technologies to be intuitive and user-friendly. The latter group was also more likely to anthropomorphize AI in their narratives. The paper thus concludes that AI literacy promotes more accurate perceptions of AI and supports a more nuanced distinction between artificial and human intelligence.

4.2.5 How Organizations Can Foster Responsible Human–AI Collaboration [Paper X]

Paper X argues that the implementation of AI in the workplace may substantially alter power relations, work structures, norms of collaboration, and responsibilities (Giermindl et al., 2022; O'Neill et al., 2022). Since the deployment of AI is influenced not only by top-down management decisions but also by employee-driven initiatives (Hammerschmidt et al., 2024; Krakowski, 2025), it spreads unevenly across organizations, making it more difficult to develop effective strategies for managing human–AI collaboration. This decentralization creates challenges in balancing the different—and sometimes conflicting—goals of stakeholders (Bonnefon et al., 2024), which may lead to unintended negative effects for both professionals and the organization. A key challenge is finding the right approach to

deploying AI in a way that aligns with diverse stakeholder interests while minimizing risks (Berente et al., 2021; Heyder et al., 2023).

Drawing on 27 interviews with professionals, **Paper X** outlines key measures for organizations to responsibly manage human–AI collaborations in response to four decision dilemmas when making strategic choices about AI deployment. First, organizations should strengthen the capabilities for reflection among professionals, potentially through an ethics committee, to vertically integrate diverse stakeholder perspectives. This addresses the tension between promoting a culture of shared responsibility and maintaining a clear allocation of individual accountabilities, particularly in collaborative tasks where responsibility gaps may arise. Second, organizations suggest that fostering moral values requires cooperation among lawyers, managers, and technicians, as well as structured dialogue to horizontally align differing moral viewpoints. This balance is critical when leveraging AI's capabilities while mitigating risks of ethical violations through, for instance, using loopholes. Third, organizations must actively embrace diversity to reconcile the tension between standardization and individualized processes, particularly in content generation tasks. Fourth, professionals should receive training on AI's limits coupled with transparent communication about its appropriate applications. Allowing trial-and-error learning is suggested to build trust in AI, but it must be balanced against the risk of overtrust, an issue most relevant in early-stage, human-like AI applications. Thus, empathetic leadership and compliance oversight are suggested as essential to mitigate friction between organizational efficiency goals and employee wellbeing. This seems particularly relevant in established processes where maximizing AI utility may overshadow personal interests. Thus, **Paper X** advances current research by identifying organizational responses to decision dilemmas in AI deployment. In doing so, it demonstrates how to promote responsible human–AI collaboration and integrate ethical considerations on a strategic level to navigate decision dilemmas.

5 Discussion and Implications

This dissertation's main objective was twofold. First, it aimed to increase our understanding of how professionals perceive responsibilities in human–AI collaborations (RQ1), and second, it identified abilities and structural conditions enabling responsible human–AI collaborations (RQ2). The following sections highlight this dissertation's key implications for research (Section 5.1) and practice (Section 5.2), and then some limitations that provide avenues for future research are outlined (Section 5.3).

5.1 Theoretical Implications

This section brings together the key findings across this dissertation's two tightly linked research questions (RQ1 and RQ2) and presents five implications for IS research, which will be detailed in the following:

- **(i)** The notions of sociomateriality allow for systematically anticipating implications of human–AI collaborations in organizations
- **(ii)** Distinguishing feeling responsible (psychological ownership) from being responsible (formal accountability) makes responsibility tensions transparent
- **(iii)** Aiming for responsible human–AI collaboration requires extending the concept of AI literacy by considering influencing factors
- **(iv)** Analyzing informal learning processes may illuminate skilling gaps through tensions between required upskilling and concurrent deskilling
- **(v)** While AI literacy can increase confidence in how AI operates, it may paradoxically reduce willingness to follow even accurate AI advice.

(i) The Notions of Sociomateriality Allow for Systematically Anticipating Implications of Human–AI Collaborations in Organizations

This dissertation advances IS research by providing a lens for systematically examining how technologies and people mutually shape each other and are entangled in organizational life (see Section 2.1). **Paper I** develops an analytical framework that integrates sociotechnical systems theory (Leonardi, 2012) with core concepts from sociomateriality (the notions of sociomateriality: *inseparability*, *performativity*, *relationality*, *practices*, and *materiality* brought forward by Jones (2014)) to render visible the cyclical interdependencies and relations linking humans (as social entities), AI (as material entities), and managerial practices (as part of the social subsystem). Responding to the call of Berente et al. (2021) to embed sociomaterial perspectives more deeply into theorizing, the framework developed in **Paper I** does not treat sociomateriality as an outcome of scholarly description, but positions sociomateriality as a normative lens for guiding reflection on responsible human–AI collaboration, a call that also more recent literature makes (e.g., Rowe et al., 2024). In doing so, **Paper I** acknowledges that AI functionality and affordances evolve dynamically within sociomaterial practices, in line with recent scholarship (e.g., Hinds & Von Krogh, 2024). In accordance with this, this dissertation outlines that, for instance, people's characteristics such as AI literacy influence their advice-taking from AI (**Paper VII**) and their AI perceptions (**Paper IX**), and individuals' morality impacts their perceived responsibilities for outcomes of human–AI collaboration (**Paper IV**).

The findings can be applied to studying real-world examples, such as the rollout of an AI recruiting system (combining e.g., résumé screening, interview scheduling, and candidate "fit" scoring) within a global HR function. A traditional approach to analyzing the changes that the introduction of this system brings might focus on time-to-hire reductions or track adoption among recruiters. Using a sociomaterial lens, this dissertation suggests foregrounding the entangled interplay of human agency, technological capabilities, and organizational practices, thereby showing that hiring outcomes do not simply *result from* the system but emerge through entanglements among recruiters, hiring managers, candidates, and organizational practices, and that they have further impacts. From this point, the AI is more than a neutral tool; for instance, it actively reshapes what counts as "merit" (e.g., which features are taken as proxies for competence) (**Paper I**), and redistributes decision rights (e.g., from hiring managers to People Analytics), impacting role responsibilities (**Paper V**). By making these interdependencies visible, the framework presented in this dissertation offers a more granular account of how AI transforms, for instance, recruiting practices and management. Moreover, it provides a conceptual basis for systematically discussing normative tensions through tensions among competing values and principles (e.g., fairness vs. efficiency, privacy vs. data utility), that may result in decision dilemmas further discussed in **Paper X**.

(ii) Distinguishing Feeling Responsible (Psychological Ownership) from Being Responsible (Formal Accountability) Makes Responsibility Tensions Transparent

In light of this dissertation's findings from experiments on responsibility attribution in joint decision-making with AI (**Paper III** and **Paper IV**) and from analyzing current philosophical discourse (**Paper II**), this dissertation encourages future research to differentiate between "being responsible" in a formal sense (linked to deontological ethics) and "feeling responsible" at a psychological level (linked to virtue ethics), required to ensure responsible human–AI collaboration (Passlack et al., 2024), since conflating the two may undermine governance. **Paper II** synthesizes the cross-disciplinary debate by examining perceptions of moral agency and responsibility in human–AI collaboration, demonstrating how these views vary with the underlying normative framework (e.g., deontological versus virtue-ethical perspectives) and serving as a theoretical foundation for future research.

By contrast, **Paper III** provides empirical findings showing that when decisions are AI-advised, professionals attribute less responsibility to themselves for failures than for successes; the same results were obtained for the reference group, where participants received advice from human consultants. Notably, this asymmetry occurs despite legal-philosophical arguments emphasizing that AI cannot be held accountable in the same way as human actors (Shneiderman, 2020). The dissertation thus exposes a responsibility tension: while someone must be responsible for negative outcomes, no entity (social or material) perceives itself as fully accountable. These findings extend to human–AI collaborations the well-known self-serving (hedonic) bias—attributing success to oneself and failure to external factors—known from human-human contexts (e.g., Chandler et al., 1990; Greene et al., 2022). This can have further implications, as the psychological experience of responsibility influences both individual well-being (Weiner, 1985) and the potential for responsible actions in organizations. Accordingly, this dissertation argues that effective governance requires analytically distinguishing feeling responsible (psychological ownership) from being responsible (formal accountability) and realigning accountability structures in a way that formal duties are supported by the conditions under which professionals actually feel responsible in practice.

(iii) Aiming for Responsible Human–AI Collaboration Requires Extending the Concept of AI Literacy by Considering Influencing Factors

Moreover, this dissertation extends the concept of AI literacy by orienting it explicitly toward *responsible* human–AI collaboration (**Paper VI**). This means considering unintended consequences such as bias-driven exclusion of minority groups (e.g., Verma, 2022), ethical considerations (further discussed in Heyder et al. (2023)), or reduced self-reflective behavior among users (e.g., Passlack, Gerholz, et al., 2025). It argues that, under conditions of imperfect AI (Lebovitz et al., 2021) and rapid technological change, personal dispositions become increasingly critical (**Paper VI**). Since AI capabilities evolve quickly, sustained self-directed learning (as personal disposition) is often the only viable means for professionals to keep pace with current system functions and limitations (Vial, 2019). Consequently, AI literacy may be framed as a set of dynamic abilities prioritizing context-sensitive responsiveness to the context of AI application, type of AI, and technological change. Furthermore, this dissertation argues that the concept of AI literacy should account for a range of influencing factors at different analysis levels that shape skill development (**Paper VI**). Notably, based on interviews, deskilling appears to be primarily influenced by micro-level factors, such as the AI system's usability (**Paper VI**) and individuals' motivational orientation (**Paper VII**). In contrast, upskilling is more frequently reported in connection with meso-level factors, such as a departmental culture that actively supports learning and development in the context of AI integration (**Paper VI**).

Likewise, the dissertation proposes understanding AI literacy as a multiliteracy influenced by adjacent literacies. Some findings indicate that ethical literacy is a prerequisite for AI literacy when the objective is responsible outcomes as opposed to those that are merely efficiency-oriented; others emphasize that technical understanding is necessary to act responsibly (**Paper VI**). The effects are mutually reinforcing since programming may accomplish a task; yet without ethical consideration, model parameters can still yield discriminatory results. The dissertation thus underscores the importance of sequencing and interdependencies among literacies when designing interventions for responsible human–AI collaboration, broadening the discourse on what constitutes AI literacy for responsible outcomes of human–AI collaborations and challenging narratives that focus predominantly on technical upskilling.

(iv) Analyzing Informal Learning Processes May Illuminate Skilling Gaps Through Tensions Between Required Upskilling and Concurrent Deskilling

Another theoretical implication of this dissertation is that studying informal learning processes not only can inform AI literacy conceptualizations but can also reveal tensions between required abilities for responsible human–AI collaboration and unintended deskilling processes.

Tensions between required skills and deskilling were linked, for instance, to critical reflection. Both the prior literature (**Paper VI**) and this dissertation's empirical findings indicate that critical thinking becomes increasingly essential in the context of collaboration with AI (**Paper VI** and **Paper VII**), especially with generative AI, which produces content based on statistical approximations and is therefore prone to hallucinations, especially when data are incomplete or lacking (Maynez et al., 2020). Based on interviews, this dissertation reveals, however, that a majority of professionals report engaging in less critical reflection when acting upon AI-generated AI suggestions, often reporting time savings as a key reason (**Paper VII**). This points to a potential skilling gap in professional practice.

An additional noteworthy finding of **Paper VII** concerns the divergent perceptions of creativity in human–AI collaboration. On the one hand, some professionals report enhanced creativity, viewing AI as a sparring partner, a source of second opinions, or a tool for brainstorming (**Paper V** and **Paper VII**). On the other hand, some interviewees perceive a decline in their own creative capacities as a result of using AI, suggesting risks of deskilling (**Paper VII**). This is supported in the study of **Paper VI**, which suggests a skilling problem, highlighting the tension between required upskilling and deskilling. In response, this dissertation highlights initial explanations for these differing experiences, which need further empirical investigations. **Paper VII** suggests that individuals with intrinsic motivation to perform good in their human–AI collaboration report a higher sense of deskilling, whereas those driven by extrinsic motivations tend to perceive greater upskilling. Thus, future research is urged to combine investigations on informal and formal learning processes, as research suggests that approximately 70–90% of competencies are best developed through learning-by-doing rather than formal training (Cerasoli et al., 2018; Ludwig, 2024).

(v) While AI Literacy Can Increase Confidence in How AI Operates, It May Paradoxically Reduce Willingness to Follow Even Accurate AI Advice

This dissertation is, to the best of my knowledge, among the first to link AI literacy with correspondence bias and attribution theory, thereby bridging psychological theory and IS research. It also advances the discourse on AI literacy by providing empirical evidence regarding its complex and sometimes paradoxical effects on advice-taking from AI. Research to date paints a mixed picture of how AI knowledge shapes reliance on AI. Some studies find that people with greater AI knowledge (a facet of AI literacy) are more likely to rely on AI by, for example, delegating tasks or taking its advice (Panigutti et al., 2022; Pinski et al., 2023). Those results, though, come from settings in which the AI's suggestions were unusually strong, which may overstate reliance. Other work shows the opposite pattern: people with stronger statistical literacy tend to trust algorithms less (Marmolejo-Ramos et al., 2025). At the same time, a limited understanding of AI can make systems seem smarter than they are, increasing advice-taking (Ehsan et al., 2021). Supported by work suggesting that lower levels of AI literacy lead to more AI advice-taking (Tully et al., 2025). Clarifying these relationships matters because generative AI systems are not perfect; in fact, they are inherently fallible (Lebovitz et al., 2021). This work sheds light on the relationship between AI literacy and AI-based advice-taking by combining quantitative experiments with qualitative analyses (**Paper VIII**).

Through experiments, this dissertation shows that individuals with higher AI literacy are less likely to rely on AI advice (**Paper VIII**). This reduced advice-taking hurt performance when the AI's recommendations were correct, but helped when its guidance was flawed. In other words, AI literacy fosters a critical stance that may lower efficiency in certain contexts but enhances resilience against imperfect outputs (**Paper VIII**). **Paper IX** and **Paper VIII** add to this by analyzing qualitative data, revealing that AI literacy influences how individuals perceive AI. Participants with higher literacy demonstrate a more balanced and differentiated view, acknowledging both opportunities and limitations of AI technologies (**Paper VIII** and **Paper IX**). By contrast, individuals with lower AI literacy levels frequently anthropomorphize AI and attribute human-like qualities to technological systems (**Paper VIII** and **Paper IX**) and show a tendency to have overly high expectations of generative AI (**Paper VIII**). These results indicate that AI literacy may not only improve the accuracy of people's perceptions of AI, but also increase skepticism in advice-taking—an effect that can reduce performance when the

AI's recommendations are, in fact, error-free (**Paper VIII**), contributing to IS research by offering a more holistic view of AI literacy and its diverse implications for collaboration with AI.

5.2 Practical Implications

This dissertation also offers implications for the management of human–AI collaborations. In doing so, it contributes to practice by clarifying how professionals may attribute responsibility in human–AI collaboration settings (RQ1) and identifying the abilities and organizational conditions that enable responsible collaboration (RQ2). The two implications below synthesize key findings across the papers, thereby providing practice recommendations:

- **(i)** Minding the responsibility gaps is a leadership and governance challenge at least as much as a technical one
- **(ii)** AI literacy ought to be treated as a toolbox, not to substitute for, but to augment reskilled professionals

(i) Minding the Responsibility Gaps Is a Leadership and Governance Challenge at Least as Much as a Technical One

In relation to RQ1—how professionals perceive responsibilities in human–AI collaboration—one key practical implication of this dissertation is that responsible human–AI collaboration may not arise solely from optimizing and explaining algorithms (linked to explainable AI (R. Dwivedi et al., 2023)), but also from stronger accountability architectures in organizations and user characteristics. Across online experiments, findings show that professionals claim more responsibility when AI-advised decisions succeed than when they fail (**Paper III**), revealing a responsibility gap. However, individuals with a stronger moral identity tend to attribute more responsibility to themselves in both success and failure, narrowing this gap (**Paper IV**). Likewise, **Paper VI** emphasizes, it is required to complement technical abilities with morality and reflective judgment in order to ensure responsible human–AI collaboration and going beyond regulatory compliance (**Paper VI**). This highlights that, for instance, foundational knowledge in statistics, combined with a reflective judgement and moral values aligned with organizational principles, are critical for long-term success (**Paper VI)**. In practical terms, technical expertise is indispensable for understanding how AI systems function and for evaluating their outputs. However, such expertise contributes to responsible AI use only when applied to reach responsible outcomes, rather than to exploit system limitations for personal advantage.

A compelling beginning approach is to align professionals' felt responsibility with formal accountability so that ownership is not only assigned but genuinely embraced. In the decision-making context, "felt responsibility" is understood as the psychological ownership that "this decision is mine," while formal accountability is the role-based obligation to stand by outcomes (**Paper III**). To avoid these divergences in practice, domain expertise must remain decisive, meaning AI may inform judgment but it cannot evaluate consequences on the professionals' behalf (**Paper VI**). Large language models can provide coherent reasons that should be treated as inputs to, not substitutes for, deliberation. This necessitates that professionals justify their decisions to skeptical stakeholders without appealing to the authority of the AI system as an advisor. For high-impact decisions, this may also include establishing clear escalation paths for ethical concerns and requiring documented rationales. Where these standards are not met, processes and capabilities may be reinforced. Moreover, the dissertation provides initial

insights into how certain role responsibilities may change through increased AI deployment (**Paper V**), which have further implications for management. To summarize, this dissertation recommends:

Recommendation 1: *Institutionalize explanatory disclosure as a standard component of decision-making, foster critical reflection among professionals, and delineate new role responsibilities where necessary.*

Further, the dissertation shows that morality is highly dependent on culture and subject views and the different ethical streams individuals follow (**Paper II**). **Paper X** demonstrates how, at times, conflicting stakeholder views can generate ethical dilemmas, such as whether to standardize processes for higher efficiency or individualize them to account for more diversity in outcomes. Organizations are, therefore, encouraged to foster ethical dialogue on the appropriate and inappropriate uses of AI across different contexts and tasks, recognizing that individuals bring varied individual views (virtue ethics) to these discussions (**Paper II**). This discourse would include discussions on what should be relevant to the decision that needs to be made and where to deploy AI (**Paper X**). For instance, in low-stakes use cases, such as drafting first-pass summaries or organizing notes, AI may accelerate work with little downside because errors are easy to catch and consequences are limited. By contrast, **Paper II** and **Paper III** suggest that in high-impact decisions such as hiring determinations and safety-critical operations, AI may inform, but human judgment is essential for weighing context, values, and the ethical implications that models cannot fully capture (**Paper II**). This approach resonates with discourse ethics, which emphasizes the consideration of multiple perspectives (Morley et al., 2021) and the pursuit of cross-cultural consensus (**Paper II**).

Management suggestions would be to foster an understanding of shared responsibility and ethical discussion among professionals (**Paper VI**), supported by prior research (e.g., Cañas, 2022) that includes the development team that programs the algorithm and therefore ensures ethics by design, but also users who use the decision support systems. While users often cannot fully understand how AI works, but developers can, developers cannot always anticipate how users may (mis)use the system (e.g., Berente et al., 2021). This underscores the need for ethical discourse in organizations at a strategic level (**Paper I**). Those may include interdisciplinary dialogue and joint initiatives with industry partners, academia, or governmental bodies, enabling organizations to embed ethical considerations more firmly into their AI strategies. Going beyond legal requirements, organizations may establish ethical principles aligned with their organizational structures that exceed existing legal requirements. This is supported by findings from qualitative interviews, in which leaders emphasized that current regulatory frameworks, often perceived as fragmented and too vague for practical application, fail to provide clear guidance on morally acceptable thresholds (**Paper X**). These insights underscore the need for more context-driven initiatives to complement regulation and address critical gaps in ethical guidance (**Paper X**). Together, these insights highlight the need for promoting discourse ethics by integrating deontological principles in the form of guidelines, while also providing space for considerations rooted in virtue ethics (**Paper VI**):

Recommendation 2: *Cultivate a shared understanding of responsibility and promote discourse ethics by embedding deontological principles into actionable guidelines, while also creating space for considerations grounded in virtue ethics.*

(ii) AI literacy Ought to Be Treated as a Toolbox, Not to Substitute for, but to Augment Reskilled Professionals

This dissertation also provides implications for management regarding which abilities become critical for responsible human–AI collaborations. It synthesizes findings from the literature (**Paper VI**), identifying skills that gain importance with AI collaborations, while also incorporating professionals' perceptions on their ability development through their AI use, in the sense of informal learning (**Paper VII**), thereby combining theoretical and practice-oriented viewpoints. Taken together, these insights offer guidance for managers and educational programs by highlighting the needs for reskilling to ensure responsible human–AI collaboration and, at the same time, mitigating deskilling.

Concerning this, a central finding is that responsible human–AI collaboration requires a role-ability fit: professionals' role responsibilities must align with their actual ability to act responsibly (**Paper VI**). For instance, in the context of a bank, having decision-making power with respect to granting loans would require being able to understand the parameters on which the AI systems calculate creditworthiness. Building on this, **Paper VI** proposes a toolbox perspective of AI literacy—a flexible and multidimensional framework that can be adapted to various professional contexts and levels of responsibility. We need to acknowledge that ability requirements are influenced by contextual factors, depending on the point of time of interaction with AI (development, deployment, etc.), as well as by the specific AI technology in use, the context of application, and so on (**Paper VI**). Furthermore, **Paper VI** suggests that learnable "hard skills," such as technical competencies that are perishable and must be continuously updated, are critical. In addition, personality dispositions such as motivation to learn, moral identity and ethical literacy, and the ability to recognize, reflect upon, and deliberate ethical dimensions of AI use, serve as more stable prerequisites for responsible AI collaboration and are less easily acquired through short-term training (**Paper VI**). Based on these findings, the third recommendation states:

Recommendation 3: *To ensure a role-ability fit, organizations should define role-specific ability requirements, consider personal disposition, and understand AI literacy as a toolbox and embed it into processes of continuous organizational learning.*

Moreover, the findings demonstrate that AI deployment within organizations is often decentralized and rarely steered exclusively by top-down managerial decisions (**Paper VII** and **Paper X**). As a result, professionals may acquire abilities through informal learning practices, such as trial-and-error, experimentation with AI applications, and peer exchange, rather than through formal workshops or structured training (Ludwig, 2024). These AI applications may foster creativity—for instance, by offering a "second opinion" or by stimulating brainstorming with novel prompts (**Paper VII**). At the same time, some findings also suggest that reliance on generative AI tools (e.g., ChatGPT) for writing tasks can encourage copy-and-paste use without sufficient critical reflection (**Paper VII**). This underscores the significant potential of generative AI for knowledge work, while simultaneously highlighting risks of deskilling. While the rapid pace of technological development makes informal processes particularly relevant, they may require organizational scaffolding.

For management, the implication is to evaluate more carefully which tasks could benefit from generative AI and which do not. In domains where critical reflection is essential—such as research (investigated in **Paper VII**)—organizations may establish clear guidelines on appropriate use, including when and how

to verify facts, and should consider mechanisms within the workflow or the AI system that explicitly prompt users to reflect critically on outputs. This outlines the need for management to shift professionals from strictly transactional tool use to iterative augmentation (e.g., prompt refinement), fostering AI literacy and thus protecting core skills (e.g., critical reflection) from deskilling.

The dissertation provides evidence that AI literacy influences AI use dynamics. While higher AI literacy enhances critical reflection, it may also reduce reliance on AI advice (**Paper VIII**). This creates a paradox: skepticism resulting from increased AI literacy can safeguard against errors, such as hallucinations, but AI aversion may prevent organizations from realizing efficiency gains from the effective deployment of AI. The findings, therefore, point to the need for a balanced approach to trust in AI. Organizations may aim to foster a culture that simultaneously encourages professionals to experiment with AI (trial-and-error learning) while cultivating critical awareness of its limitations (**Paper VI** and **Paper X**). Understanding responsible human–AI collaboration as a joint optimization of people and systems, organizations should aim to train professionals (AI literacy, learning-by-doing) while engineering usability, transparency, and safe defaults into AI. This would require calibrating the depth of explanations, available controls, and usability to stakeholder groups at different AI literacy levels (e.g., leaders vs. employees; teachers vs. students). This leads to the fourth recommendation:

Recommendation 4: *Design mechanisms that allow for hands-on experimentation and non-formal learning, complementing structured training where professionals surface and share their AI use cases, aiming for augmentation rather than substitution of professional expertise.*

5.3 Limitations & Future Research Avenues

As with any scholarly work, this dissertation has limitations that stem from the methods applied and from the focus and scope of the works included. These limitations create valuable opportunities for further inquiry along five avenues for future research:

- **Avenue 1**. Multi-stakeholder perceptions of responsibility
- **Avenue 2**. Addressing biases in responsibility attribution
- **Avenue 3**. Factors influencing AI literacy development
- **Avenue 4**. The strategic relevance of non-formal learning
- **Avenue 5**. Long-term effects of human–AI collaboration and real-world validation

Avenues 1 and 2 are linked primarily to RQ1, focusing on understanding responsibility perceptions in human–AI collaborations. Avenues 3 and 4 address RQ2, examining the abilities and structural conditions required to develop AI literacy for responsible human–AI collaboration. Avenue 5 concerns the limitations that come with the methods applied.

Avenue 1. Multi-Stakeholder Perceptions of Responsibility

This dissertation advances IS research by examining how responsibility is perceived in human–AI collaboration, but several limitations qualify its contributions. The analysis concentrates on perceived moral responsibility and does not develop a comprehensive conceptual or legal account of responsibility. While **Paper II** synthesizes discussions from ethics, IS, and business, it stops short of a holistic philosophical treatment of responsibility in human–AI collaboration, and liability issues in the sense of legal responsibility are addressed only superficially. The experimental investigations are descriptive,

analyzing reported self-attribution of moral responsibility in joint decision making (**Paper III** and **Paper IV**). Moreover, the empirical focus is primarily on professionals such as managers (**Paper III** and **Paper IV**) and programmers (**Paper VIII**) receiving AI-based advice for solving tasks, limiting the applicability of findings to other professional groups and contexts. The dissertation draws on data only from German and English speakers, which may further limit the generalizability of its findings. As **Paper X** illustrates, cultural norms and regulatory frameworks may shape how responsibility is perceived and enacted. These boundaries constrain generalizability but also delineate clear possibilities for future research.

Future research could extend beyond managerial decision-making to ethically charged, high-stakes contexts—most notably autonomous driving, where outcomes may have life-or-death implications and effects may therefore be stronger or qualitatively different (Passlack et al., 2023; Rowe et al., 2024). Rowe et al.'s (2024) literature review on responsibility in the context of autonomous driving offers a useful point of departure: while roughly 80% of the studies surveyed address liability issues, moral responsibility in autonomous driving remains comparatively underexamined and warrants deeper inquiry (Rowe et al., 2024). Responsibility in practice is also shaped by a diverse set of stakeholders and cultural settings (**Paper X**). Comparative, cross-cultural designs would better capture differing moral norms and regulatory regimes and assess their impact on responsibilities in human–AI collaboration. Future work is encouraged to examine both self-attributions of responsibility by those directly collaborating with AI and responsibility attributions by external observers (e.g., building on the work such as of Kegel & Ghanem (2024)), and should broaden the actor set to include developers as well as policymakers, customers, and business partners. This would further illuminate how moral responsibility is distributed in sociotechnical systems and the extent to which AI advice may inadvertently legitimize misbehavior.

Avenue 2. Addressing Biases in Responsibility Attribution

This dissertation shows that biases originally found in human–human settings also manifest in human–AI collaboration. **Paper VII** demonstrates that professionals claim less responsibility for outcomes when they relied on AI advice and the outcome was unsuccessful than when successful. This pattern indicates that the hedonic bias extends to human–AI interactions, thereby creating responsibility gaps—particularly salient because AI systems cannot be held legally accountable (Constantinescu et al., 2021; Dyrkolbotn et al., 2018). **Paper VIII** further shows that when generative AI provides bad programming advice, individuals tend to attribute the failure to the model's inherent disposition, such as a bad model, rather than to situational factors such as incomplete task information. This is consistent with the correspondence bias, which states that individuals tend to attribute behavior to dispositional rather than situational factors (Gilbert & Malone, 1995). This dispositional framing also reduced subsequent reliance on AI advice, even when later recommendations were accurate (**Paper VIII**). These findings are limited to a very specific context and require further explanation of the attribution logics applied to AI and how to address such biases.

Promising directions for future research include systematic investigation of other cognitive biases that may surface in human–AI collaboration. These may include, for example, automation bias, that is, the tendency to lean too heavily on AI recommendations, and confirmation bias, that is, the inclination to accept advice that fits one's pre-existing beliefs (Leffrang et al., forthcoming). Furthermore, future research is encouraged to develop mitigation strategies. While **Paper VIII** provides initial propositions

on how AI literacy can mitigate correspondence bias and **Paper IV** outlines that morality influences self-attribution for decision-making, thereby mitigating hedonic bias, it further indicates the importance of individual traits. Additional work could assess the extent of this effect and compare it with technology-centric interventions. In particular, design patterns of AI systems warrant closer study: anthropomorphic cues can impact the extent to which systems are blamed (Pavone et al., 2023); trust in a system is influenced by how predictable its behavior is (Karnouskos, 2021; Rowe et al., 2024); and perceived intentionality of observed behavior influences evaluations of moral responsibility (Kegel & Ghanem, 2024; Papagiannidis et al., 2025). Consequently, another avenue for future research concerns how to mitigate biases by identifying and evaluating technology design factors such as transparency and predictability, and understanding how users navigate, adapt to, and compensate for opacity would offer actionable guidance for managing responsibility gaps in human–AI collaborations.

Avenue 3. Factors Influencing AI Literacy Development

Another avenue for future research concerns the identified skilling gaps. While critical reflection is perceived as important, some findings suggest a decline in critical reflection within human–AI collaborations (**Paper VII**). This calls for tracking early-warning indicators alongside AI literacy growth, such as declining performance on cornerstone tasks without AI support (Strich et al., 2021), reduced ability to explain or justify decisions (suggested in **Paper VII**), overreliance on default model outputs (investigated in **Paper VIII**), and decreased diversity in idea generation (**Paper VI**). Notably, findings point to heterogeneous perceptions of creativity: some users report gains, others losses (**Paper VII** and **Paper X**). Recent literature mirrors this divergence, indicating that generative AI can enhance user creativity (Ali Elfa & Dawood, 2023; Doshi & Hauser, 2024) while potentially reducing the individuality and novelty of outputs (Doshi & Hauser, 2024). Clarifying these mixed effects is beyond the scope of this dissertation but remains an important task for future work. Initial findings highlight the role individual motivation may have in perceived skill development (**Paper II**); research suggests that people who are less skilled may benefit more from generative AI ideas, and it can even harm the creative process of higher-skilled individuals (Yin et al., 2024). While there may be other factors, there is limited evidence on how personality traits shape informal learning (Tannenbaum & Wolfson, 2022).

Prior research further indicates that AI's impact on skill development is strongly conditioned by technological design (Bankins et al., 2024). Moreover, individuals may adapt certain behavior from AI through increased collaborations with AI, referred to as the "behavior adoption effect" (Heyder et al., 2023). To maintain innovativeness in generating ideas and avoid decreased critical reflection, future research is also invited to systematically examine potential influencing factors that may impact deskilling and upskilling of AI literacy. This may include not only individual factors, but also design features that may hinder or foster critical reflection, such as through prompts like "tell me first what you think." Deeper empirical investigations are needed to trace how such effects would manifest across settings and technologies.

Avenue 4. The Strategic Relevance of Non-Formal Learning

This dissertation integrates two complementary strands of evidence to examine changes in ability requirements for responsible human–AI collaboration; both, though, rely on interview-based perceptions and do not directly measure or quantify these abilities. In **Paper VI**, practitioner interviews identify the abilities most relevant for collaborating with AI responsibly, and **Paper VII** examines how

professionals perceive their own skill development when using AI in practice. The latter line of inquiry is likely to become increasingly important for two reasons. Many AI tools—especially generative systems such as ChatGPT, Claude, and GitHub Copilot—are adopted informally by employees rather than through formal organizational rollouts (e.g., Krakowski, 2025). At the same time, research consistently finds that most capability growth occurs "on the job," through learning-by-doing rather than through formal courses (Cerasoli et al., 2018; e.g., Ludwig, 2024). Consistent with prior work (e.g., Vial, 2019), this dissertation also shows that continuous learning itself is becoming a critical ability as technologies evolve rapidly (**Paper VI**). Together, these points suggest that non-formal learning processes are gaining strategic importance and warrant sustained attention from both researchers and practitioners.

Existing conceptualizations of AI literacy—often derived from qualitative literature analyses (e.g., Ng et al., 2021) and expert interviews (e.g., Heyder & Posegga, 2021)—usefully inform training design. A focus on technical abilities, ethical literacy, or critical reflection, however, may struggle to keep pace with the rapid introduction of new systems. Personalized learning offers one response that is becoming increasingly relevant (Hwang et al., 2025), but it remains largely a formal learning approach that needs to be conceptualized. By contrast, non-formal learning is typically task-driven: individuals aim to solve work problems, with learning occurring as a byproduct rather than an explicit goal (Werquin, 2010).

Examine how organizations can systematically cultivate non-formal learning to build AI literacy as part of a people-development strategy is a promising avenue for future research. While this may appear paradoxical—embedding non-formal learning in organizational strategy can render it "formal"—the intent here follows Tannenbaum and Wolf's (2022) view that organizations can deliberately support and scaffold informal learning while preserving its emergent, task-driven character. This includes management considerations regarding which structures and practices are most effective in fostering learning in professionals' daily practice. As Tannenbaum and Wolfson (2022) summarize, these may include identifying effective structures and practices such as supportive cultures and job designs, psychological safety fostered by managers, peer support, and recognition systems. Future research could investigate how to move from course-based formal education programs to continuous learning environments where non-formal learning is the primary engine. This could also include developing ways to measure growth in AI literacy as it happens within the sociomaterial entanglement, for instance, through longitudinal self- and supervisor ratings that link learning activities to observable improvements in work.

Avenue 5. Long-Term Effects of Human–AI Collaboration and Real-World Validation

Overall, the findings of this dissertation are shaped by the methods employed and the cross-sectional nature of the studies included. Vignette experiments, while useful for identifying causal relationships, cannot fully capture the complexity of real-world settings (Eifler & Petzold, 2019); moreover, participants often exhibit social desirability bias when reporting on moral behavior. The literature reviews, though systematic and transparent, are constrained by database selection, time windows, screening criteria for quality and relevance, and a disciplinary focus on IS as well as ethics and management. Interviews can provide rich insights but may have limited generalizability (Gibbs, 2018; Sarker et al., 2013). Qualitative text analyses are susceptible to our interpretive judgments, even when mitigated through extensive discussion among researchers and the use of a shared codebook, following the guidelines of Mayring (2014). To address these constraints, future research should adopt mixed-

methods, which "offer the opportunity to address exploratory and confirmatory questions of IS research within the same research inquiry" (Reis et al., 2022, p. 557; Venkatesh et al., 2016). Longitudinal designs would better capture the dynamics of AI literacy development over time and reveal how responsibility attributions may shift as AI behavior adapts. Field studies would enhance validity by observing whether the dissertation's findings are valid in real-world settings.

6 Conclusion

In conclusion, this dissertation has been guided by two interrelated research questions: How do professionals perceive responsibility in human–AI collaborations (RQ1), and which abilities and structural conditions enable such collaborations to be responsible (RQ2)? By addressing them in tandem, this dissertation provides an initial understanding of what responsible human–AI collaboration may look like. Grounded in sociotechnical systems theory and sociomateriality, the analysis adopts a relational view on human–AI collaboration and treats responsibility as an emergent property of entangled professionals, AI technologies, roles, and routines rather than as a fixed attribute of any single entity.

The findings of this dissertation contribute to IS research at the intersection of management, education, ethics, and psychology in five complementary ways. First, drawing on sociomaterial notions by Jones (2014), the work shows how research and organizations may systematically analyze the implications of human–AI collaboration (**Paper I**), making, for instance, redistributions of responsibility visible before they entrench into organizational norms. Second, it clarifies how responsibilities are perceived and negotiated in human–AI collaborations (**Paper II**) and how role responsibilities may shift with increased AI utilization (**Paper V**), and it renders responsibility tensions analytically tractable by distinguishing feeling responsible (psychological ownership) from formal accountability (**Papers III**), thereby also indicating how these tensions may be mitigated by professionals' moral identity (**Paper IV**). Third, it extends the concept of AI literacy beyond individual skills to include the conditions that determine whether those skills translate into responsible collaboration with AI, including influencing factors at the micro and meso-level (**Paper VI**). Fourth, it shows how informal learning processes surface skilling gaps due to discrepancies between required reskilling and deskilling processes, pinpointing where managerial support is required (**Paper VII**). Fifth, it uncovers AI literacy's effect on AI use: as understanding of AI mechanics deepens with increased AI literacy, both confidence in one's understanding of system behavior and more realistic perceptions of AI increase (**Paper IX**), yet utilization of AI recommendations may decline (**Paper VIII**). This has direct implications for training, AI deployment, and expectation management, where this work provides insights into possible strategies for how to respond to decision dilemmas (**Paper X**).

Together, these contributions can offer theoretical grounding for future IS research toward better understanding and managing responsible human–AI collaboration, as well as practical guidance for organizational management seeking to translate AI literacy and governance into everyday, responsible human–AI collaboration.

One lesson learned stands out from my four years of research: tolerance of ambiguity is indispensable. Cultivating the ability to navigate uncertainty, complexity, and indeterminacy with composure has proven essential—and will remain so as technological and social shifts continue to accelerate.

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

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, *25*(1), 107–136. https://doi.org/10.2307/3250961

Ali Elfa, M. A., & Dawood, M. E. T. (2023). Using artificial intelligence for enhancing human creativity. *Journal of Art, Design and Music*, *2*(2), 3.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), Article 3. https://doi.org/10.1080/09528130050111428

Angelopoulos, S., Bendoly, E., Fransoo, J., Hoberg, K., Ou, C., & Tenhiälä, A. (2023). Digital transformation in operations management: Fundamental change through agency reversal. *Journal of Operations Management*, *69*(6), 876–889. https://doi.org/10.1002/joom.1271

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction*, *4*(2), Article 2. https://doi.org/10.3390/make4020026

Anton, E., Behne, A., & Teuteberg, F. (2020). The Humans Behind Artificial IN^{TEL}LIGENCE – An Operationalisation of AI Competencies. In *European Conference on Information Systems (ECIS)*.

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., & McCutcheon, L. E. (2019). How Serious Is the `Carelessness' Problem on Mechanical Turk? *International Journal of Social Research Methodology*, *22*(5), 441–449. https://doi.org/10.1080/13645579.2018.1563966

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, *6*(3), 128–138. https://doi.org/10.1027/1614-2241/a000014

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 Quartlery*, *45*(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Baker, M. (2000). Writing a Literature Review. *The Marketing Review*, *1*, 219–247. https://doi.org/10.1362/1469347002529189

Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., & Beekhuyzen, J. (2015). Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support. *Communications of the Association for Information Systems*, *37*(1). https://doi.org/10.17705/1CAIS.03708

Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, *185*(4), Article 4. https://doi.org/10.1007/s10551-023-05339-7

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, *45*(2), 159–182. https://doi.org/10.1002/job.2735

Barnes, S. B. (1999). *Understanding Agency: Social Theory and Responsible Action*. 1–176.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2024). *Lme4: Linear Mixed-Effects Models Using "Eigen" and S4*. https://cran.r-project.org/package=lme4

Bawden, D. (2008). Origins and Concepts of Digital Literacy. *Digital Literacies: Concepts, Policies and Practices*, *30*, 17–32.

Beauchamp, T. L. (2003). The nature of applied ethics. *A Companion to Applied Ethics*, *26*, 1–17.

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), 10.

Benlemlih, M., & Bitar, M. (2018). Corporate social responsibility and investment efficiency. *Journal of Business Ethics*, *148*, 647–671.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3).

Black, J. E., & Reynolds, W. M. (2016). Development, reliability, and validity of the Moral Identity Questionnaire. *Personality and Individual Differences*, *97*, 120–129. https://doi.org/10.1016/j.paid.2016.03.041

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

Boos, D., Guenter, H., Grote, G., & Kinder, K. (2013). Controllable accountabilities: The Internet of Things and its challenges for organisations. *Behaviour & Information Technology*, *32*(5), 449–467. https://doi.org/10.1080/0144929X.2012.674157

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly*, *25*(1), 1–16. https://doi.org/10.2307/3250956

Bradley, J. H., Paul, R., & Seeman, E. (2006). Analyzing the structure of expert knowledge. *Information & Management*, *43*(1), 77–91. https://doi.org/10.1016/j.im.2004.11.009

Bravo, E. (2015). Deskilling, up-skilling or reskilling? The effects of automation in information systems context. *Twenty-First Americas Conference on Information Systems,*. https://scholar.archive.org/work/wudaea4vmrfezldpwnnmyrosgu/access/wayback/https://aisel.aisnet.org/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1121&context=amcis2015

Brendel, A. B., Hildebrandt, F., Dennis, A. R., & Riquel, J. (2023). The paradoxical role of humanness in aggression toward conversational agents. *Journal of Management Information Systems*, *40*(3), 883–913. https://doi.org/10.1080/07421222.2023.2229127

Brendel, A. B., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. *Sustainability*, *13*(4), Article 4. https://doi.org/10.3390/su13041974

Bright, D. S., Winn, B. A., & Kanov, J. (2014). Reconsidering virtue: Differences of perspective in virtue ethics and the positive social sciences. *Journal of Business Ethics*, *119*, 445–460.

Burton-Jones, A., Butler, B. S., Scott, S., & Xu, S. X. (2021). Next-generation information systems theorizing: A call to action. *MIS Quarterly*, *45*(1), 301–314.

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology*, *13*, 836650.

Carroll, N., Holmström, J., Umeå University, Stahl, B. C., University of Nottingham, Fabian, N. E., & University of Groningen. (2024). Navigating the Utopia and Dystopia Perspectives of Artificial Intelligence. *Communications of the Association for Information Systems*, *55*, 854–874. https://doi.org/10.17705/1CAIS.05533

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, *1*(2), 74–78.

Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The sociomateriality of information systems. *MIS Quarterly*, *38*(3), 809–830.

Cerasoli, C. P., Alliger, G. M., Donsbach, J. S., Mathieu, J. E., Tannenbaum, S. I., & Orvis, K. A. (2018). Antecedents and Outcomes of Informal Learning Behaviors: A Meta-Analysis. *Journal of Business and Psychology*, *33*(2), 203–230. https://doi.org/10.1007/s10869-017-9492-y

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. https://doi.org/10.1109/TEM.2021.3138503

Chandler, T., Seibel, C., & Spies, C. (1990). Attributions and Affects Related to Actual Vs Perceived Success in an Actual Achievement Context. *Psychological Reports*, *66*(3), 971–976. https://doi.org/10.2466/PR0.66.3.971-976

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems*, *39*(4), 969–1005. https://doi.org/10.1080/07421222.2022.2127441

Chen, W., & Hirschheim, R. (2004). *A paradigmatic and methodological examination of information systems research from 1991 to 2001*. *14*(3), 197–235. https://doi.org/10.1111/j.1365-2575.2004.00173.x

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Chuang, S. (2020). An empirical study of displaceable job skills in the age of robots. *European Journal of Training and Development*, *45*(6/7), 617–632. https://doi.org/10.1108/EJTD-10-2019-0183

Churchill, L. R. (1999). Are We Professionals? A Critical Look at the Social Role of Bioethicists. *Daedalus*, *128*(4), 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*, *X.X*(1), 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Conboy, K. (2019). Being Promethean. *European Journal of Information Systems*, *28*(2), 119–125. https://doi.org/10.1080/0960085X.2019.1586189

Conboy, K., Fitzgerald, G., & Mathiassen, L. (2012). Qualitative methods research in information systems: Motivations, themes, and contributions. *European Journal of Information Systems*, *21*(2), 113–118. https://doi.org/10.1057/ejis.2011.57

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), Article 4. https://doi.org/10.1007/s10676-021-09616-9

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*(1), 205316801882217. https://doi.org/10.1177/2053168018822174

Cramton, C. D. (2002). Attribution in distributed work groups. *Distributed Work*, 191–212. https://doi.org/10.7551/mitpress/2464.003.0014

Creswell, J. W., & Creswell, D. J. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). SAGE.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Daily, M., Oulasvirta, A., & Rekimoto, J. (2017). Technology for human augmentation. *Computer*, *50*(2), 12–15.

Dalkir, K. (2013). *Knowledge Management in Theory and Practice*. Routledge.

Dean, J. W., & Sharfman, M. P. (1996). Does Decision Process Matter? A Study of Strategic Decision-Making Effectiveness. *The Academy of Management Journal*, *39*(2), 368–396. https://doi.org/10.2307/256784

Dennis, A. R., & Valacich, J. S. (2001). Conducting Experimental Research in Information Systems. *Communications of the Association for Information Systems*, *7*(5), 1–41. https://doi.org/10.17705/1cais.00705

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

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

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv*, 1–47. https://doi.org/10.48550/arXiv.2107.13509

Eifler, S., & Petzold, K. (2019). Validity Aspects of Vignette Experiments: Expected "What-If" Differences Between Reports of Behavioral Intentions and Actual Behavior. In *Experimental Methods in Survey Research* (pp. 393–416). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119083771.ch20

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science*, *331*(6018), 721–725. https://doi.org/10.1126/science.1201765

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferguson, L. W. (1960). Ability, interest, and aptitude. *Journal of Applied Psychology*, *44*(2), 126–131.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Finch, J. (1987). The Vignette Technique in Survey Research. *Sociology*, *21*(1), 105–114. https://doi.org/10.1177/0038038587021001008

Fincham, F., & Jaspars, J. (1980). Attribution of Responsibility: From Man the Scientist to Man As Lawyer. In *Advances in Experimental Social Psychology* (Vol. 13, pp. 81–138). https://doi.org/10.1016/S0065-2601(08)60131-8

Fink, A. (2019). *Conducting Research Literature Reviews: From the Internet to Paper*. Sage publications: Thousand Oaks, CA, USA.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814594

Flathmann, C., Schelble, B. G., & McNeese, N. J. (2021). Fostering Human-Agent Team Leadership by Leveraging Human Teaming Principles. *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582649

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2021). How to Design AI for Social Good: Seven Essential Factors. In L. Floridi (Ed.), *Ethics, Governance, and Policies in Artificial Intelligence* (Vol. 144, pp. 125–151). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_9

Frazer, E. (2006). Max Weber on Ethics and Politics. *Journal of International Political Theory*, *2*(1), 19–37. https://doi.org/10.1177/1743453X0600200103

Frimer, J. A., & Walker, L. J. (2009). Reconciling the self and morality: An empirical model of moral centrality development. *Developmental Psychology*, *45*(6), 1669.

Garrison, D. R., Cleveland-Innes, M., Koole, M., & Kappelman, J. (2006). Revisiting methodological issues in transcript analysis: Negotiated coding and reliability. *The Internet and Higher Education*, *9*(1), 1–8. https://doi.org/10.1016/j.iheduc.2005.11.001

Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, *36*(1), Article 1. https://doi.org/10.1080/08839514.2021.2009222

Gewirth, A. (1960). Meta-ethics and normative ethics. *Mind*, *69*(274), 187–205.

Gibbs, G. R. (2018). *Analyzing qualitative data* (2nd ed.). SAGE Publications. https://www.torrossa.com/gs/resourceProxy?an=5017782&publisher=FZ7200

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

Gilbert, D., & Malone, P. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38. https://doi.org/10.1037/0033-2909.117.1.21

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

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), Article 4. https://doi.org/10.1007/s10551-022-05056-7

Greene, P., Latham, A. J., Miller, K., & Norton, J. (2022). Capacity for simulation and mitigation drives hedonic and non-hedonic time biases. *Philosophical Psychology*, *35*(2), 226–252. https://doi.org/10.1080/09515089.2021.1960299

Grodal, S., Krabbe, A. D., & Chang-Zunino, M. (2023). The Evolution of Technology. *Academy of Management Annals*, *17*(1), 141–180. https://doi.org/10.5465/annals.2021.0086

Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems*, *35*(4), 1092–1120. https://doi.org/10.1080/07421222.2018.1523534

Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology*, *20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307–320. https://doi.org/10.1007/s10676-017-9428-2

Hager, W. (2006). Die Fallibilität empirischer Daten und die Notwendigkeit der Kontrolle der Wahrscheinlichkeiten falscher Entscheidungen. *Zeitschrift für Psychologie / Journal of Psychology*, *214*(1), 10–23. https://doi.org/10.1026/0044-3409.214.1.10

Hakli, R., & Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist*, *102*(2), 259–275. https://doi.org/10.1093/monist/onz009

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/

Harder Fischer, L., Nicolajsen, H. W., Marttila, S.-M., & Sandbukt, S. (2024). Crafting Meaningful Generative AI-Enabled Knowledge Work. *Proceedings/European Conference on Information Systems (ECIS)*, *2024*, 10. https://aisel.aisnet.org/ecis2024/track01_peoplefirst/track01_peoplefirst/10/?trk=public_post_comment-text

Hardy, S. A., & Carlo, G. (2011). Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action?: Moral Identity Development. *Child Development Perspectives*, *5*(3), 212–218. https://doi.org/10.1111/j.1750-8606.2011.00189.x

Harvey, J. H., Weary, G., & Harris, B. (1981). *Perspectives on Attributional Processes*. W.C. Brown Company.

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

Heider, F. (1958). *The Psychology of Interpersonal Relations* (1st ed.). Wiley.

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. *International Conference on Information Systems (ICIS) 2021*, *9*.

Hinds, P., & Von Krogh, G. (2024). Generative AI, Emerging Technology, and Organizing: Towards a theory of progressive encapsulation. *Organization Theory*, *5*(4), 26317877241293478. https://doi.org/10.1177/26317877241293478

Hong, J. W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Horváth, I. (2022). AI in interpreting: Ethical considerations. *Across Languages and Cultures*, *23*(1), 1–13. https://doi.org/10.1556/084.2022.00108

Hwang, G.-J., Yilmaz, R., & Wu, J.-Y. (2025). Generative Artificial Intelligence-enhanced Adaptive/Personalized Learning. *Computers & Education - Call for Papers*. https://www.sciencedirect.com/special-issue/319587/generative-artificial-intelligence-enhanced-adaptivepersonalized-learning

Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., & Sycara, K. (2018). Transparency and Explanation in Deep Reinforcement Learning Neural Networks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 144–150. https://doi.org/10.1145/3278721.3278776

Jaiswal, A., Arun, C. J., & Varma, A. (2022). Rebooting employees: Upskilling for artificial intelligence in multinational corporations. *The International Journal of Human Resource Management*, *33*(6), 1179–1208. https://doi.org/10.1080/09585192.2021.1891114

Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of Basic and Clinical Pharmacy*, *5*(4), 87.

Jeong, S., Han, S. J., Lee, J., Sunalai, S., & Yoon, S. W. (2018). Integrative Literature Review on Informal Learning: Antecedents, Conceptualizations, and Future Directions. *Human Resource Development Review*, *17*(2), 128–152. https://doi.org/10.1177/1534484318772242

Jones, M. (2014). A matter of life and death: Exploring Conceptualizations of Sociomateriality in the Context of Critical Care. *MIS Quarterly*, *38*(3), 895-A6. https://doi.org/10.25300/MISQ/2014/38.3.12

Jurczyszyn, E. (2024). Exploring the Dynamic Landscape of Formal, Informal and Non-Formal Learning. *International Journal of Pedagogy, Innovation and New Technologies*, *11*(1), 53–59.

Kagan, S. (2018). *Normative Ethics* (1st ed.). Routledge.

Kamar, E. (2016). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. *IJCAI*, 4070–4073. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/hi.pdf

Karnouskos, S. (2021). The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cognition, Technology & Work*, *23*(4), 659–667. https://doi.org/10.1007/s10111-020-00649-6

Kegel, 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*. https://scholarspace.manoa.hawaii.edu/items/f5eca98b-8adc-418a-8881-985001b4c43f

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2022). *AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers*. http://arxiv.org/abs/2207.01493

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

King, W. R., & He, J. (2005). Understanding the Role and Methods of Meta-Analysis in IS Research. *Communications of the Association for Information Systems*, *16*(1). https://doi.org/10.17705/1CAIS.01632

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. *2*.

Koukouvinou, P., & Holmström, J. (2022). AI management beyond the narratives of dystopian nightmares and utopian dreams: A systematic review and synthesis of the literature. *European Conference on Information Systems (ECIS)*. https://aisel.aisnet.org/ecis2022_rp/137/

Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization*, *35*(1), 100560.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice and using software*. SAGE Publications Ltd.

Lapsley, D. K., & Narvaez, D. (2004). A Social–Cognitive Approach to the Moral Personality. In *Moral Development, Self, and Identity*. Psychology Press.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, 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), 1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Leffrang, D., Passlack, N., Müller, O., & Posegga, O. (forthcoming). Beneficial Mistrust in Generative AI: The Role of AI Literacy in Handling Bad Advice. *BISE*.

Lemmon, E. J. (1962). Moral dilemmas. *The Philosophical Review*, *71*(2), 139–158.

Leonardi, P. M. (2012). *Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How are They Related? Do We Need Them?* (SSRN Scholarly Paper 2129878). https://doi.org/10.2139/ssrn.2129878

Li, O. (2021). Problems with “Friendly AI.” *Ethics and Information Technology*, *23*(3), Article 3. https://doi.org/10.1007/s10676-021-09595-x

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

Ludwig, A. (2024). *Organisationales Lernen und Digitalisierung: Lernen von und in Organisationen unter dem Einfluss der digitalen Transformation*. Springer Fachmedien. https://doi.org/10.1007/978-3-658-44153-1

Mabaso, B. A. (2021). Computationally rational agents can be moral agents. *Ethics and Information Technology*, *23*(2), Article 2. 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

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. https://doi.org/10.1007/s12599-019-00600-8

Marmolejo-Ramos, F., Marrone, R., Korolkiewicz, M., Gabriel, F., Siemens, G., Joksimovic, S., Yamada, Y., Mori, Y., Rahwan, T., & Sahakyan, M. (2025). Factors influencing trust in algorithmic decision-making: An indirect scenario-based experiment. *Frontiers in Artificial Intelligence*, *7*, 1465605.

Marshall, C., & Rossman, G. B. (2006). *Designing Qualitative Research*. SAGE Publications.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), Article 4. 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), Article 2. https://doi.org/10.1007/s11023-021-09556-9

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, *19*(4). https://aisel.aisnet.org/misqe/vol19/iss4/6

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

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), 147:1-147:19. https://doi.org/10.1145/3415218

McGuire, J., & De Cremer, D. (2023). Algorithms, leadership, and morality: Why a mere human effect drives the preference for human over algorithmic leadership. *AI and Ethics*, *3*(2), 601–618. https://doi.org/10.1007/s43681-022-00192-2

McHugh, M. L. (2012). Lessons in biostatistics: Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 359–372. https://doi.org/10.1201/b16720-37

Meuser, M., & Nagel, U. (1989). Experteninterviews-vielfach erprobt, wenig bedacht: Ein Beitrag zur qualitativen Methodendiskussion. *Nr. 6 Des Sonderforschungsbereichs 186 Der Uni Bremen ‚Statuspassagen Und Risikolagen Im Lebensverlauf. Bremen.* https://www.ssoar.info/ssoar/handle/document/5773

Mikalef, P., 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

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE.

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

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Moore, D., Haines, K., Drudik, J., Arter, Z., & Foley, S. (2020). Upskill/Backfill Model of Career Pathways Advancement: The Nebraska Vocational Rehabilitation Approach. *Journal of Applied Rehabilitation Counseling*. https://doi.org/10.1891/JARC-D-20-00002

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, *40*(2), 120–123.

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

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., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*(100041), 1–11. https://doi.org/10.1016/j.caeai.2021.100041

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), Article 2. https://doi.org/10.14746/fped.2017.6.2.20

Noy, I. Y., Shinar, D., & Horrey, W. J. (2018). Automated driving: Safety blind spots. *Safety Science*, *102*, 68–78.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*. https://hal.science/hal-01574600

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *65*(5), 604–938. https://doi.org/10.1177/0018720820960865

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.

Orlikowski, W. J., & Barley, S. R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? *MIS Quarterly*, 145–165.

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research*, *2*(1). https://doi.org/10.1287/isre.2.1.1

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

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), Article 1. https://doi.org/10.1017/S0963180123000087

Panigutti, C., Beretta, A., Giannotti, F., & Pedreschi, D. (2022). Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. *Proceedings of the Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3491102.3502104

Pansoni, S., Tiribelli, S., Paolanti, M., Di Stefano, F., Malinverni, E. S., Frontoni, E., & Giovanola, B. (2023). Artificial Intelligence and Cultural Heritage: Design and Assessment of AN Ethical Framework. *29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage. Humanities and Digital Technologies for Shaping the Future," 48*, 1149–1155. https://u-pad.unimc.it/handle/11393/315590

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885.

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., Tate, M., Johnstone, D., & Kitsiou, S. (2016). Contextualizing the twin concepts of systematicity and transparency in information systems literature reviews. *European Journal of Information Systems*, *25*(6), 493–508. https://doi.org/10.1057/s41303-016-0020-3

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

Passlack, N., Gerholz, K.-H., & Schlottmann, P. (2025, January 7). Reskilling Me Softly: Perceived Changes in Students' Skilling When Using GenAI in Academic Research Projects. *Proceedings of the 58th Hawaii International Conference on System Sciences*. https://hdl.handle.net/10125/109440

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., Hammerschmidt, T., & Posegga, O. (2025). With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI? *Information Systems Frontiers*, 1–30. https://doi.org/10.1007/s10796-025-10648-5

Passlack, N., Heyder, T., Klemm, F., & Posegga, O. (2023). How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success. *ICIS 2023 Proceedings*. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/15

Pavone, G., Meyer-Waarden, L., & Munzel, A. (2023). Rage Against the Machine: Experimental Insights into Customers' Negative Emotional Responses, Attributions of Responsibility, and Coping Strategies in Artificial Intelligence–Based Service Failures. *Journal of Interactive Marketing*, *58*(1), 52–71. https://doi.org/10.1177/10949968221134492

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, *31*(4), 623–656.

Pinski, M., Adam, M., & Benlian, A. (2023). AI Knowledge: Improving AI Delegation through Human Enablement. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3580794

Pinski, M., & Benlian, A. (2023). AI Literacy—Towards Measuring Human Competency in Artificial Intelligence. *Proceedings of the 56th Hawaii International Conference on System Sciences*.

Pregenzer, M., Remus, U., & Wiener, M. (2021). Algorithms in the Driver's Seat: Explaining WorkersÄ Reactions to Algorithmic Control. *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*, *30*, 1–15. https://doi.org/10.1080/08839514.2016.1229919

Rafner, J., Dellermann, D., Hjorth, A., Veraszto, D., Kampf, C., MacKay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: A case for hybrid intelligence. *Morals & Machines*, *1*(2), 24–39.

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

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.

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: Assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems*, *31*(3), 364–387.

Reinhardt, U. E. (1992). Reflections on the Meaning of Efficiency: Can Efficiency Be Separated from Equity? *Yale Law & Policy Review*, *10*(2), 302–315.

Reis, L., Maier, C., & Weitzel, T. (2022). Mixed-Methods in Information Systems Research: Status Quo, Core Concepts, and Future Research Implications. *Communications of the Association for Information Systems*, *51*(1), 556–579. https://doi.org/10.5194/gi-2016-9-SC1

Retkowsky, J., Hafermalz, E., & Huysman, M. (2024). Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons*, *67*(5), 511–523. https://doi.org/10.1016/j.bushor.2024.04.009

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

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). http://arxiv.org/abs/2204.07612

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976

Rowe, F. (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems*, *23*(3), 241–255. https://doi.org/10.1057/ejis.2014.7

Rowe, F., Jeanneret Medina, M., Journé, B., Coëtard, E., & Myers, M. (2024). Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems. *Journal of Information Technology*, *39*(3), 587–615. https://doi.org/10.1177/02683962231207108

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline. *MIS Quarterly*, *43*(3), 695-A5.

Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. *MIS Quarterly*, *37*(4), iii–xviii.

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 1/2). *Journal of the Association for Information Systems*, *19*(8), 752–774.

Sawant, R., Thomas, B., & Kadlag, S. (2022). Reskilling and Upskilling: To Stay Relevant in Today's Industry. *International Review of Business and Economics*, *7*(1). https://doi.org/10.56902/IRBE.2022.7.1.4

Schlagwein, D., Currie, W., Leimeister, J. M., & Willcocks, L. (2025). Digital futures: Definition (what), importance (why) and methods (how). *Journal of Information Technology*, *40*(1), 2–8. https://doi.org/10.1177/02683962241301544

Schlimbach, R., & Khosrawi-Rad, B. (2022). Towards Ethical Design Features for Pedagogical Conversational Agents. *Americas Conference on Information Systems (AMCIS)*. https://aisel.aisnet.org/amcis2022/sig_service/sig_service/1/

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

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

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), Article 2. https://doi.org/10.2458/v7i2.20321

Stets, J. E., Burke, P. J., Serpe, R. T., & Stryker, R. (2020). Getting Identity Theory (IT) Right. In S. R. Thye & E. J. Lawler (Eds.), *Advances in Group Processes* (Vol. 37, p. 0). Emerald Publishing Limited. https://doi.org/10.1108/S0882-614520200000037007

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

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*, *22*(2). https://doi.org/10.17705/1jais.00663

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

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), Article 2. https://doi.org/10.1080/14494035.2021.1928377

Tannenbaum, S. I., & Wolfson, M. A. (2022). Informal (Field-Based) Learning. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(Volume 9, 2022), 391–414. https://doi.org/10.1146/annurev-orgpsych-012420-083050

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

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. Humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*, 1962. https://doi.org/10.3389/fpsyg.2017.01962

Trier, M., Kundisch, D., Beverungen, D., Müller, O., Schryen, G., Mirbabaie, M., & Trang, S. (2023). Digital Responsibility: A Multilevel Framework for Responsible Digitalization. *Business & Information Systems Engineering*, *65*(4), 463–474. https://doi.org/10.1007/s12599-023-00822-x

Tully, S. M., Longoni, C., & Appel, G. (2025). Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity. *Journal of Marketing*, *89*(5), 1–20. https://doi.org/10.1177/00222429251314491

Van De Poel, I. (2011). The Relation Between Forward-Looking and Backward-Looking Responsibility. In N. A. Vincent, I. Van De Poel, & J. Van Den Hoven (Eds.), *Moral Responsibility* (Vol. 27, pp. 37–52). Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_3

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

Venkatesh, V., Brown, S. A., & Sullivan, Y. (2016). Guidelines for conducting mixed-methods research: An extension and illustration. *Journal of the AIS*, *17*(7), 435–495.

Verma, P. (2022). These robots were trained on AI. They became racist and sexist. *The Washington Post,*. https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, *28*(2), 18–144. https://doi.org/10.1016/j.jsis.2019.01.003

Vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *ECIS 2009 Proceedings*. https://aisel.aisnet.org/ecis2009/161

Wahl, M. F., & Prause, G. (2013). Toward understanding resources, competencies, and capabilities: Business model generation approach. *Entrepreneurship and Sustainability Issues*, *1*(2), 67–80. https://doi.org/10.9770/jesi.2013.1.2(1)

Wang, B., Rau, Pei-Luen Patrick, & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 1–14. https://doi.org/10.1080/0144929X.2022.2072768

Weaver, G. R., & Trevino, L. K. (1994). Normative and empirical business ethics: Separation, marriage of convenience, or marriage of necessity? *Business Ethics Quarterly*, 129–143.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, *71*(1), 3–25. https://doi.org/10.1037/0022-0663.71.1.3

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Werquin, P. (2010). Recognising Non-Formal and Informal Learning Outcomes, Policies and Practices: Outcomes, Policies and Practices. *OECD Publishing.* https://www.oecd.org/en/publications/recognising-non-formal-and-informal-learning_9789264063853-en.html

Yamaguchi, S. (2018). Changes in returns to task-specific skills and gender wage gap. *Journal of Human Resources*, *53*(1), 32–70.

Yin, M., Jiang, S., & Niu, X. (2024). Can AI really help? The double-edged sword effect of AI assistant on employees' innovation behavior. *Computers in Human Behavior*, *150*, 107987.

Zamani, E., & Vannini, S. (2025). Digital policy narratives: Addressing grand challenges or exacerbating digital inequalities? *Information Technology for Development*, 1–20.

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, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Sciences*, *13*(2), Article 2. https://doi.org/10.3390/bs13020181

Appendix

Appendix I. Papers Included in the Cumulative Dissertation

Table I lists all papers included in this cumulative dissertation, which are presented sequentially in the following pages.

Table I. Overview of Papers Included in the Cumulative Dissertation

Paper	Title
I	Heyder, T., **Passlack, N.**, & Posegga, O. (2023). Ethical management of human–AI interaction: Theory development review. *The Journal of Strategic Information Systems (JSIS)*, *32*(3), 101772.
II	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)*, 1–87.
III	**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 (BIT)*, 1–30.
IV	**Passlack, N.**, Hammerschmidt, T., & Posegga, O. (2025). How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice. *American Conference on Information Systems 2025 Proceedings (AMCIS)*. 4.
V	**Passlack, N.** (Working Paper, unpublished). From Teachers to Cheerleaders? Teachers' Role Responsibilities in the Age of Generative AI. Previously passed the 1st revision round at *Journal of Information Technology (JIT).*
VI	**Passlack, N.**, Hammerschmidt, T., & Posegga, O. (2025). With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?. *Information Systems Frontiers (ISF)*.
VII	**Passlack, N.**, Gerholz, K. H., & Schlottmann, P. (2025). Reskilling Me Softly: Perceived Changes in Students' Skilling When Using GenAI in Academic Research Projects. *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*.
VIII	Leffrang, D., **Passlack, N.**, Müller, O., & Posegga, O. (2026). Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice (No. 136). *Business and Information Systems Engineering (BISE)*.
IX	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).*
X	Heyder, T., **Passlack, N.**, & Posegga, O. (Working Paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment. Previously progressed to a 2nd revision round at *The Journal of Strategic Information Systems (JSIS).*

Paper I

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human–AI interaction: Theory development review. *The Journal of Strategic Information Systems (JSIS)*, 32(3), 101772.

https://doi.org/10.1016/j.jsis.2023.101772

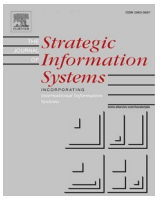
Journal of Strategic Information Systems 32 (2023) 101772



Contents lists available at ScienceDirect

Journal of Strategic Information Systems

journal homepage: www.elsevier.com/locate/jsis





Ethical management of human-AI interaction: Theory development review

Teresa Heyder, Nina Passlack, Oliver Posegga *

Information Systems and Social Networks, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Ethics
Human-AI interaction
Theoretical review
Sociomateriality

ABSTRACT

AI-based technologies have changed the nature of the symbiosis between humans and AI, and so strategic management of human-AI interaction in organizations requires deeper ethical considerations. Aligning AI with human values requires a systematic understanding of the ethical management of human-AI interaction. We conduct a theoretical review, from a sociotechnical perspective, and analyze ethical management of human-AI interaction through the lens of sociomateriality. Our systematic approach helps explain and clarify the interdependencies between two ethical perspectives – duty and virtue ethics – in sociotechnical systems. We also provide a theoretical framework that leads to seven avenues for future research.

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 pervasiveness, 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

* Corresponding author.
E-mail addresses: teresa.heyder@uni-bamberg.de (T. Heyder), nina.passlack@uni-bamberg.de (N. Passlack), oliver.posegga@uni-bamberg.de (O. Posegga).

https://doi.org/10.1016/j.jsis.2023.101772

Available online 3 August 2023
0963-8687/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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 phenomen[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.

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.).

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.

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.

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*

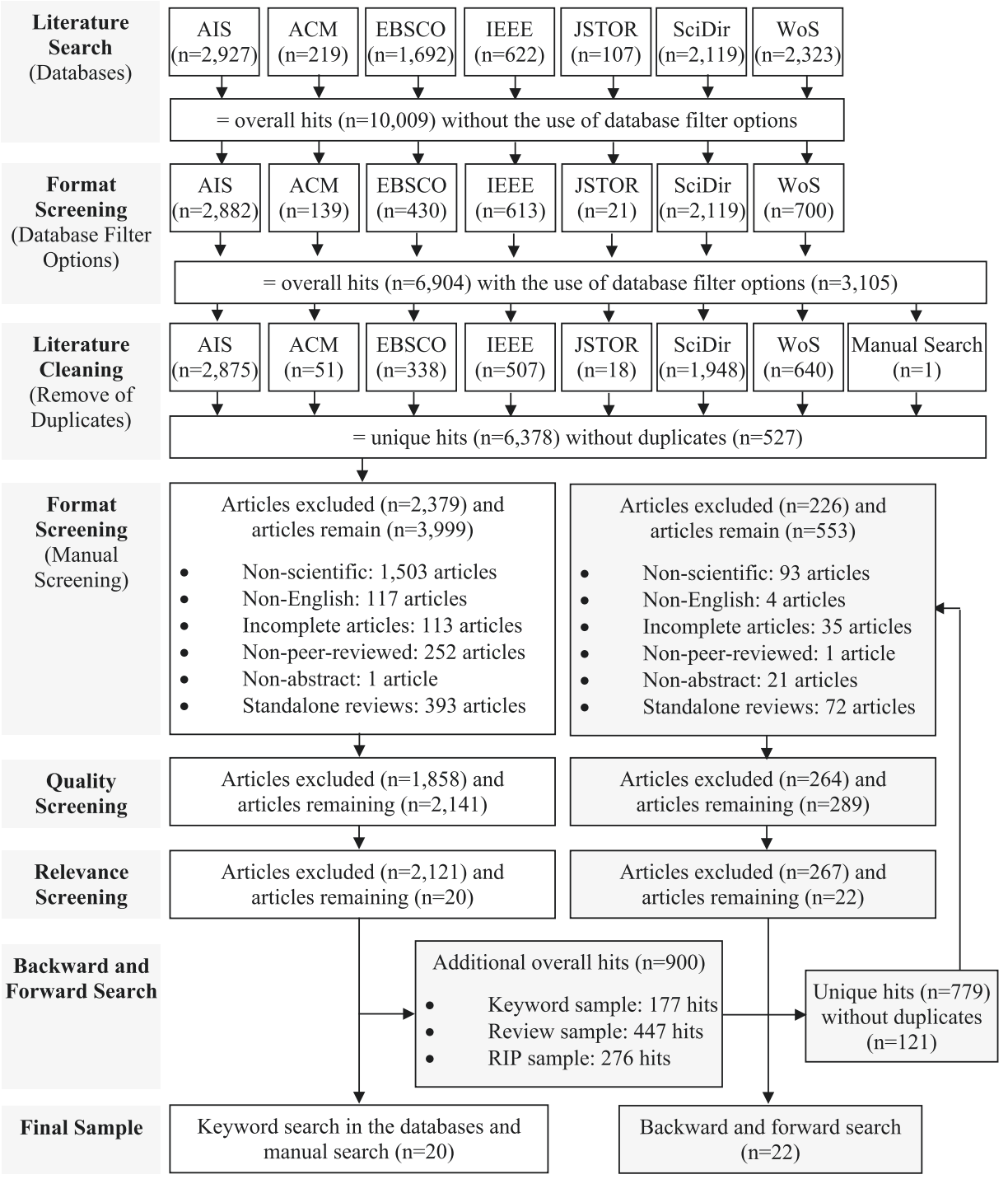


Fig. 1. Search Process, based on Templier and Paré (2018).

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,

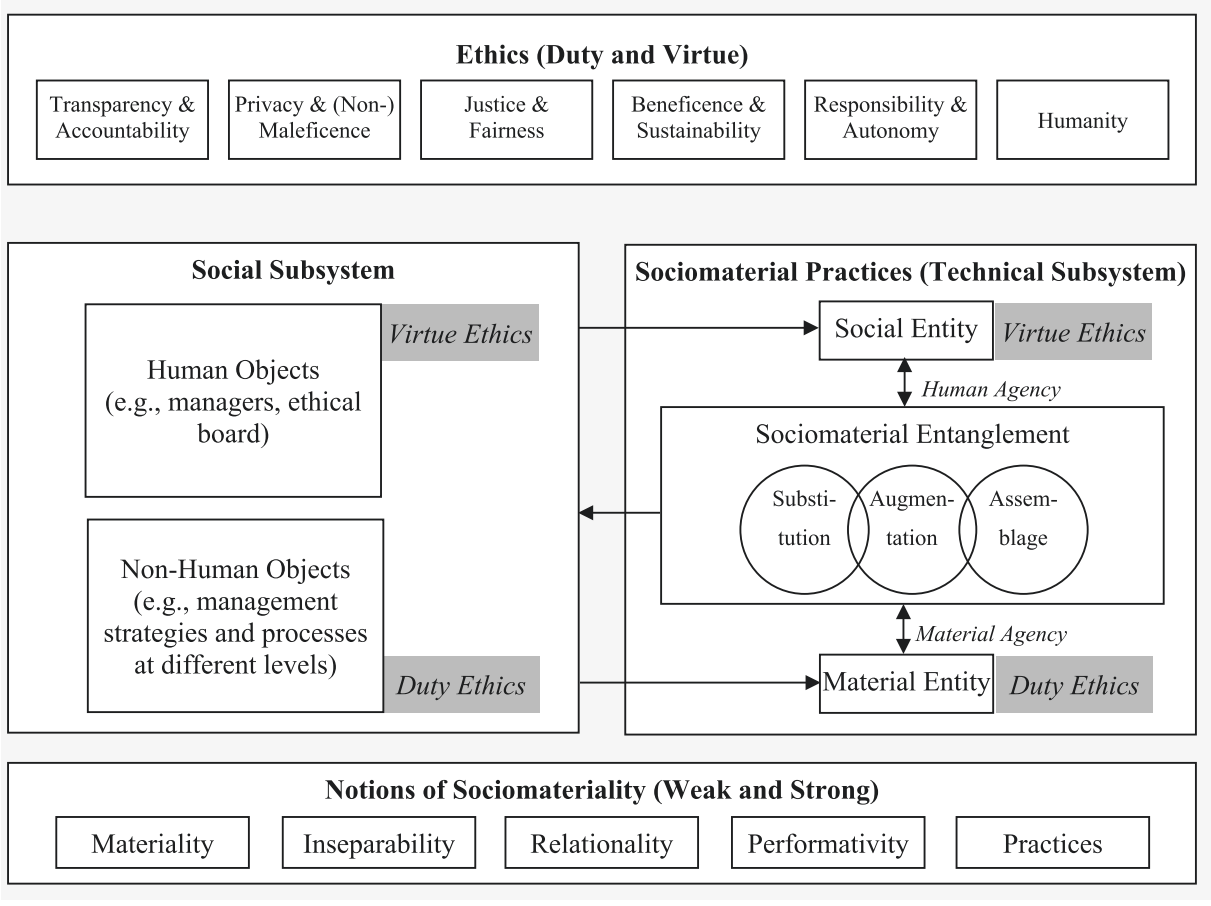


Fig. 2. Conceptual Framework of Ethical Management of Human-AI Interaction.

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.

Fig. 1 is an overview of the search process based on the steps suggested by 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.

Conceptual framework

The following pages explain the three main constructs (ethics, human-AI interaction, and strategic management) and their sub-constructs 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 Fig. 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).

As Fig. 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 Fig. 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 Fig. 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.

How the types of Human-AI interaction refer to sociomaterial entanglement

As Fig. 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.

How ethics is considered as part of both the social subsystem and sociomaterial practices

As Fig. 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.

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 Fig. 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

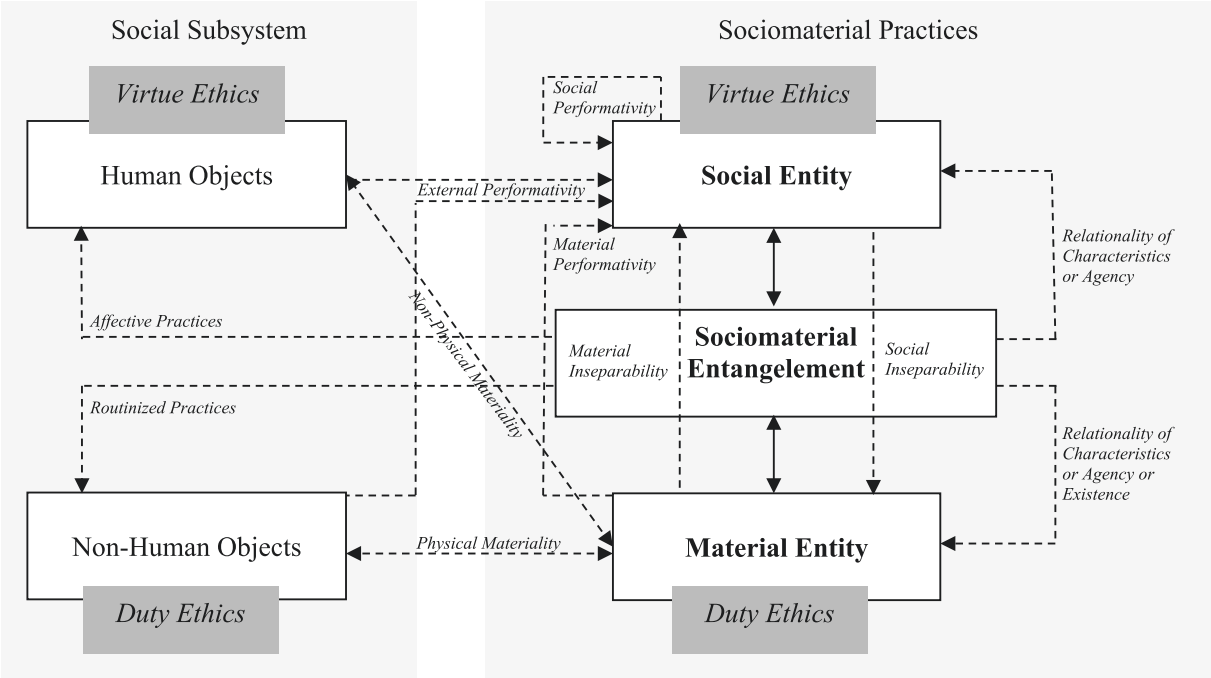


Fig. 3. Notions of Sociomateriality Determining Ethical Management in Sociotechnical Systems.

management of human-AI interaction by explaining the relationships and interdependencies between all constructs and their sub-constructs we introduced in this stage through the five notions of sociomateriality.

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.

Fig. 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.

Although Fig. 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 Fig. 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*). Pregenzer 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*

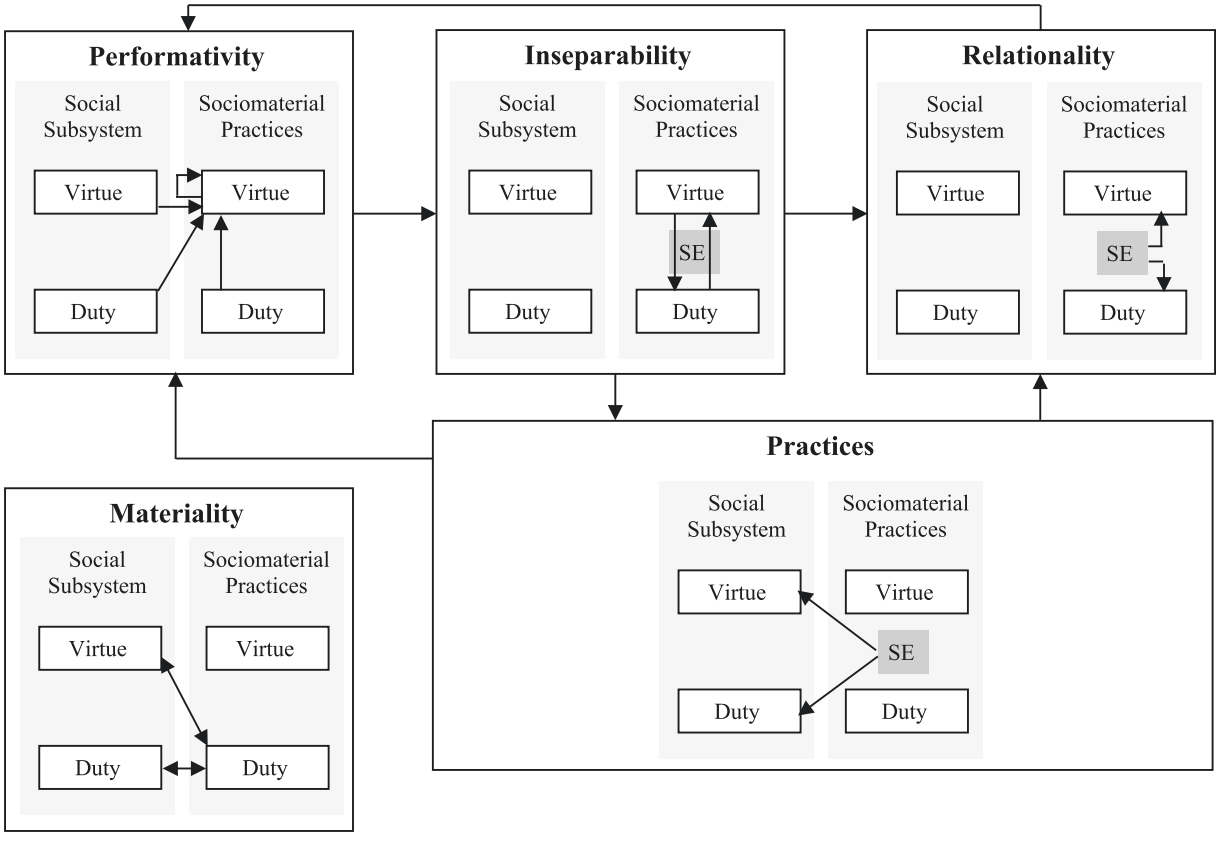


Fig. 4. The Sociomaterial Cycles.

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. Fig. 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:*

In the following, we explain the relationships between the notions illustrated in Fig. 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 adaptions 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.

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.

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 socio-technical 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
Toward a Research Agenda for Virtue Ethics in Sociotechnical System Theory.

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 is an overview of possible research regarding *virtue ethics* in the context of sociomateriality.

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" (Figueroa-Armijos et al., 2022, p. 12) and various tasks for which AI is used that "impact ethical perceptions" (Figueroa-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.

Table 2
Toward a Research Agenda for the Notions of Sociomateriality.

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)? (*physical materiality*) • To what extent does AI ethics affect the ethical values (*virtue ethics*) on a collective level (e.g., organizational culture)? (*non-physical materiality*) • To what extent does AI ethics affect the ethics of other technological objects (e.g., design principles of machine ethics)? (*digital materiality*)
Inseparability	• How can we align AI values (*duty ethics*) with human values (*virtue ethics*)? (*social inseparability*) • How can we measure the alignment between human values and AI values? (*material inseparability*)
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? (*relationality of individuals' characteristics* and *agency*) • To what extent are positive or negative changes in the existence, characteristics, or agency of individuals a result of human-AI interaction? (*relationality of individuals' characteristics* and *agency*)
Performativity	• How can training influence human agency? (*external performativity*) • How does the transparency of AI affect human agency? (*material performativity*) • How do humans build perceptions of AI? (*social performativity*)
Practices	• How do emotions arising from human-AI interaction affect an individual's *virtue ethics*? (*affective practices*) • How does *virtue ethics* of individuals differ depending on how they are used for interacting with AI? (*routinized practices*)

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; Pregenzer 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).

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 (Klumpp 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.

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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Notions of weak and strong sociomateriality

In Table A1, we adopt and extend toward our research context the *weak* and *strong* notions of sociomateriality as defined by Jones (2014).

Table A1

Notions of Weak and Strong Sociomateriality, based on 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 *strong* and *weak* (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

Relationality describes how sociomaterial entities (social and material entities) are "being produced through relation" (Jones, 2014, 898). Compared to inseparability, *relationality* 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), *relationality* 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

Performativity 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

Materiality 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, *materiality* 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. *Materiality* also comprises human objects of organizational artifacts, such as culture, language, leadership, or the management board, given that *materiality* 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

Practices 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).

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 B1.

Table B1

Quality Criteria for Framework Development.

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 *integrating type* 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 *conceptual stage* (in comparison to the *review stage*).

(*continued on next page*)

Table B1 (*continued*)

Quality criteria for frameworks aggregated and adapted from Schwarz et al. (2007, pp. 42–43)	Author Considerations
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 *weak* and *strong* 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 *conceptual stage* as well as in the *review stage* 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 *extend strategy* for theory development of Burton-Jones et al. (2021) we used our framework that reflects well-established theories on ethics and socio-materiality (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).

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 C1:

Table C1
Leading Journals and Conferences of the IS Domain and Management and Organization Science Domain.

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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)
DESRIST (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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)

(*continued on next page*)

Table C1 (*continued*)

Leading Outlet Title	Rank	Database
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)

WoS = Web of Science, SciDir = Science Direct

Leading conferences and journals of ethics domain were categorized according to different rankings given in Table C2. 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.

Table C2
Leading Journals and Conferences of 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)

WoS = Web of Science, SciDir = Science Direct

Appendix D. Manual query

The following *five* leading journals and conferences in Table D1 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 D1 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.

Table D1
Manual Search Results.

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

Appendix E. 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 E1.

Table E1
Keyword-Based Search Strategy.

Keyword-Strings	Human-AI Interaction	AND	Strategic Management	AND	Ethical Principles
keywords	substitution OR augmentation OR assemblage		"corporate strategy" OR "business strategy" OR "functional strategy" OR "planning process*" OR "strategic process*" OR "tactical process*" OR "operational process*"		transparency OR accountability OR 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***

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 F1. 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 F1 only refers to the keyword-search strategy in the databases and the manual search and not to the backward and forward searches.

Table F1
Format Screening.

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.	Through filter options: **(total: 0 articles)** No filter available Through manual screening: **(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.	Through filter options: **(total: 0 articles)** No filter available Through manual screening: **(total: 113 articles)** ACM: 3 articles, AIS: 83 articles, EBSCO: 4 articles, IEEE: 4

(*continued on next page*)

Table F1 (*continued*)

Exclusion Criteria	Exclusion Description	Number of Eliminated Articles
		articles, JSTOR: 1 article, ScienceDirect: 5 articles, Web of Science: 13 articles
Non-peer-reviewed articles (n = 669 articles)	Articles not being peer-reviewed.	Through filter options: **(total: 417 articles)** ACM: 13 articles, AIS: 4 articles, EBSCO: 392 articles, IEEE: 8 articles Through manual screening: **(total: 252 articles)** 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.	Through filter options: **(total: 385 articles)** EBSCO: 3425 articles, IEEE: 1 article, JSTOR: 37 articles, Web of Science: 2 articles Through manual screening: **(total: 117 articles)** 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.	Through filter options: **(total: 2,303 articles)** ACM: 63 articles, AIS: 45 articles, EBSCO: 525 articles, JSTOR: 49 articles, Web of Science: 1,621 articles Through manual screening: **(total: 1 article)** AIS: 1 article
Review articles (n = 393 articles)	Standalone literature review papers are excluded.	Through filter options: **(total: 0 articles)** No filter available Through manual screening: **(total: 393 articles)** ACM: 1 article, AIS: 222 articles, EBSCO: 30 articles, IEEE: 19 articles, JSTOR: 2 articles, ScienceDirect: 51 articles, Web of Science: 68 articles
Total: 5,484 articles	*Through filter options:* ***3, 105 articles*** *Through manual screening:* ***2,379 articles***	

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 G1 gives an overview of the outlets being excluded.

Table G1
Quality Screening.

Ranking Criteria	Examples of Excluded Outlets	Number of Excluded Outlets
Journal articles not fulfilling the above-mentioned criteria from the Harzing, 2020 journal list (67th 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** Distribution: 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	Conference on Emerging Technologies & Factory Automation (ETFA), International Conference on Arts and Culture, International Conference on Multi Agent Systems, International Conference on	**Total: 201 articles** Distribution: ACM: 3 articles, AIS: 88 articles, EBSCO: 16 articles, IEEE: 2 articles, ScienceDirect: 85 articles, Web of Science: 7

(*continued on next page*)

Table G1 (*continued*)

Ranking Criteria	Examples of Excluded Outlets	Number of Excluded Outlets
ranking of at least D or higher in the VHB-conference ranking list (VHB, 2011).	Technical Debt, Wuhan International Conference on e-Business, etc.	

Appendix H. Relevance screening

To be considered relevant, the articles must meet all three criteria given in Table H1. 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.

Table H1
Relevance Screening.

Nr.	Relevance Criteria	Number of Excluded Articles
	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** Distribution: 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** Distribution: AIS: 6 articles, EBSCO: 1 article, IEEE: 1 article, ScienceDirect: 2 articles
	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** Distribution: 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** Distribution: AIS: 1 article, ScienceDirect: 4 articles
	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** Distribution: 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** Distribution: AIS: 13 articles, EBSCO: 2 articles, ScienceDirect: 17 articles, Web of Science: 10 articles
1, 2 & 3	Combination of all criteria	**Total: 8 articles** Distribution: ScienceDirect: 6 article, Web of Science: 2 articles

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 I1 for an overview:

Table I1
Backward and Forward Search.

Search Type	Round	Keyword Sample Nbr. of Articles	Screening	Review Sample Nbr. of Articles	Screening	RIP Sample Nbr. of Articles	Screening
Back-ward	1st 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	1st 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
1st 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	2nd 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	2nd round	• 5 articles • Thereof, 0 duplicates	• Excluded format: • 0 articles • Excluded quality: • 3 articles • Excluded relevance: • 1 article • 1 article remain	• 108 articles • Thereof, 14 duplicates	• Excluded format: • 26 articles • Excluded quality: • 34 articles • Excluded relevance: • 31 articles • 3 articles remain	• 7 articles • Thereof, 1 duplicate	• Excluded format: • 1 article • Excluded quality: • 4 articles • Excluded relevance: • 0 articles • 1 article remain
2nd 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)	
Back-ward	3rd 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
For-ward	3rd round	-	-	• 10 articles • Thereof, 5 duplicates	• Excluded format: • 1 article	• 3 articles • Thereof, 1 duplicate	• Excluded format: • 1 article

(*continued on next page*)

Table I1 (*continued*)

Search Type	Round	Keyword Sample Nbr. of Articles	Screening	Review Sample Nbr. of Articles	Screening	RIP Sample Nbr. of Articles	Screening
					• Excluded quality: • 1 article • Excluded relevance: • 3 articles • 0 articles remain		• Excluded quality: • 1 article • Excluded relevance: • 0 articles • 0 articles remain
3rd 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)	

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 J1:

Table J1
Final Sample.

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 2nd 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 1st 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 1st 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 2nd 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 1st 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 2nd round)
10	Fuegener et al.	2021	Cognitive challenges in human-AI collaboration: investigating the path towards productive delegation	Information Systems Research	RIP sample (backward 1st round)
11	Grønsund & Aanestad	2022	Augmenting the algorithm: emerging human-in-the-loop work configurations	The Journal of Strategic Information Systems	Review sample (backward 1st 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 1st round)
14	Hoeddinghaus et al.	20	The automation of leadership functions: Would people trust decision algorithms?	Computers in Human Behavior	Review sample (backward 1st round)

(*continued on next page*)

Table J1 (*continued*)

Nbr.	Autor(s)	Year	Title	Outlet	Source
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 1st 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 1st 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 1st 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 2nd round)
23	Lee	2018	Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management	Big Data & Society	Keyword sample (forward 2nd 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 1st 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 2nd 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 1st 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 1st 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 1st 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 1st round)
37	Shin & Park	2019	Role of fairness, accountability, and transparency in algorithmic affordance	Computers in Human Behavior	Keyword sample (backward 1st 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

Appendix K. Meta Information

Table K1 provides an overview of the *meta information* of the *final sample*:

Table K1
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

(*continued on next page*)

Table K1 (*continued*)

Meta Information	Codes	Number of Articles	Article Nbr.
	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

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 L1 provides a snapshot of the deductive coding scheme.

Table L1
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: • Interdependencies 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)

Appendix M. Types of Human-AI interaction

The three types of human-AI interaction following Rai et al. (2019) are given in Table M1:

Table M1
Types of Human-AI Interaction.

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) • "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)	• 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)
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) • *Machine augmentation*: 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) • *Human augmentation*: 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) • *Human augmentation* (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)

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 N1:

Table N1
Principles of Business Ethics.

Ethical principles	Characteristics	Authors
Transparency & understandability (distinctiveness, visibility, accuracy)	• Understandable *management processes* (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)	e.g., Goodpaster, 1991; Hasnas, 2013; Winstanley et al., 1996

(*continued on next page*)

Table N1 (*continued*)

Ethical principles	Characteristics	Authors
Responsibility & leadership (authority, legitimacy, agreement, validity, obligations, credibility, control, deregulation)	• Enhance human well-being (Hasnas, 2013) • Strengthen social, ethical, and environmental credibility (Dando and Swift, 2003) • 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

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 N2. It should be noted that we chose our own classification based on the most common AI principles named by the authors.

Table N2
Principles of AI Ethics.

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 *human agency* and the *material agency* (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)

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 N3:

Table N3
Categories of Human-AI Ethics.

Categories Human-AI Ethics	Characteristics
Transparency & accountability (accuracy, explainability, explicability, distinctiveness, instrumentality, interpretability, intelligibility, reliability, replicability, understandability, visibility, verifiability)	*Duty ethics*: 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. *Virtue ethics:* 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)	*Duty ethics*: principles that prevent personal privacy and security of humans during human-AI interaction and avoid misuse of AI through humans and vice versa. *Virtue ethics:* 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)	*Duty ethics:* 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. *Virtue ethics*: 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)	*Duty ethics*: principles that aim for beneficial human-AI interactions with a long-term positive impact. *Virtue ethics*: 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)	*Duty ethics*: 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. *Virtue ethics*: 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))	*Duty ethics*: 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. *Virtue ethics:* 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.

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 O1).

Table O1
Levels of Strategic Management.

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	

We further identified four processes of strategic management given in Table O2:

Table O2
Processes of Strategic Management.

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	

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.

Acikgoz, Y., Davison, K.H., Compagnone, M., Laske, M., 2020. Justice perceptions of artificial intelligence in selection. Int. J. Select. Assess. (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., Fourney, 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 Quart. Executive (MISQE) 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 Quart. (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. Inf. Syst. Front. 24 (3), 857–875. 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 J. Women Cult. Soc. 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, Technol. Soc. 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 Quart. Execut. (MISQE) 19, 9–21. https://doi.org/10.2139/ssrn.3741983.

Berente, N., Gu, B., Recker, J., Santhanam, R., 2021. Managing Artificial Intelligence. MIS Quart. (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. Technol. Soc. 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. J. Inform. Technol. (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. Corp. Gov. 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.M. (Eds.), The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, pp. 316–334.

Bowie, N.E., 1998. A kantian theory of meaningful work. J. Bus. 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 Quart. (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 Trans. Human-Comput. 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 Sociomaterialty of Information Systems: Current Status, Future Directions. MIS Quart. (MISQ) 38 (3), 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. Educ. Psychol. Meas. 20 (1), 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. J. Bus. 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. Int. J. Commun. 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. Int. J. Inform. Manage. (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. J. Bus. 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 Manage. Rev. 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. Bus. Inform. Syst. Eng. (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. Int. J. Inf. Manag. 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), Kopenhagen, 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. J. 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. Inf. Organ. 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. J. Bus. 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., 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. Mind. Mach. 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. Inform. Syst. Res. (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, 4th ed. Routledge, New York, pp. 1–31.
Gandz, J., Hayes, N., 1988. Teaching Business Ethics. J. Bus. Ethics (JBE) 7 (9), 657–669.
Gasser, U., Almeida, V., 2017. A layered model for AI governance. IEEE Internet Comput. 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. Eur. J. Inform. Syst. (EJIS) 31, 410–435. https://doi.org/10.1080/0960085X.2021.1927213.
Goodpaster, K.E., 1991. Business Ethics and stakeholder analysis. Bus. Ethics Quart. (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. J. Bus. 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. J. Strateg. Inf. Syst. 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.
Hamilton, C., Swart, W., Stokes, G.M., 2021. Developing a measure of social, ethical, and legal content for intelligent cognitive assistants. J. Strateg. Innov. Sustain. 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 Sytems (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. J. Bus. 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 Syst. J. 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? Comput. Hum. Behav. 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. Comput. Hum. Behav. 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. Hum. Resour. Manage. 26, 313–330. https://doi.org/10.1002/hrm.3930260302.
Islam, G., Greenwood, M., 2021. Reconnecting to the social in Business Ethics. J. Bus. 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. Int. J. Hum Comput Stud. 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. Nat. Mach. Intell. 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2.
Jones, C., Parker, M., ten Bos, R., 2005. For Business Ethics. Routledge, London, UK, p. 223.
Jones, M., 2014. A matter of life and death: exploring conceptualizations of sociomateriality in the context of critical care. MIS Quart. (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. J. Bus. 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. Int. J. Log. Res. Appl. 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. J. Bus. Logist. (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 Lead (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. J. Assoc. Inform. Syst. (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, p. 301.
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.
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., Moody, G.D., Gaskin, J., Galletta, D.F., Humpherys, S.L., 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 (4), 993–1012. https://doi.org/10.25300/MISQ/2013/37.4.01.
MacInnis, D.J., 2011. A framework for conceptual contributions in marketing. J. Market. (JM) 75, 136–154. https://doi.org/10.1509/jmkg.75.4.136.
MacKenzie, S.B., 2003. The Dangers of Poor Construct Conceptualization. J. Acad. Mark. Sci. 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. Bus. Inform. Syst. Eng. (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. Int. J. 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. J. Strateg. Inf. Syst. 30 (3), 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. J. Assoc. Inform. Syst. (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., Haimerl, 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 Quart. 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. Eur. J. Inform. Syst. (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. Electron. Mark. 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. Harv. Bus. Rev.
Mutch, A., 2013. Sociomateriality — Taking the wrong turning? Inf. Organ. 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. Comput. Educ.: Artif. Intell. 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. Res.-Technol. Management (RTM) 60, 26–30. https://doi.org/10.1080/08956308.2017.1255051.
Orlikowski, W.J., 2007. Sociomaterial practices: Exploring technology at work. Organ. Stud. (OS) 28, 1435–1448. https://doi.org/10.1177/0170840607081138.
Orlikowski, W.J., 2010. The sociomateriality of organisational life: considering technology in management research. Camb. J. Econ. (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. Acad. Manage. Ann. (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., Ashrafian, H., 2019. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int. J. Med. Rob. Comput. Assisted Surg. 15 (1), 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. Inf. Manag. 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. J. Mass Commun. 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. Int. J. Manpower (IJM) 20, 297–308. https://doi.org/10.1108/01437729910286083.
Peppard, J., Ward, J., 2004. Beyond strategic information systems: towards an IS capability. J. Strateg. Inf. Syst. 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. Inf. Syst. Front. 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. Int. J. Hum Comput Stud. 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. Acad. Manag. Rev. 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. Psychol. Mark. 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. Soc 58 (3), 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. Eur. J. 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 Neurosci. 11, 120–127. https://doi.org/10.1080/21507740.2020.1740354.
Ritson, N., 2008. Strategic Management, 2nd ed. bookboon.com, p. 117 pp..
Rivard, S., 2021. Theory building is neither an art nor a science: It is a craft. J. Inform. Technol. (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.
Sauppé, 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 J. Law Technol. 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. Bus. Soc. Rev. 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 (3), 29–50.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., de Freitas, N., 2016. Taking the human out of the loop: A review of bayesian optimization. Proc. 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. Comput. Hum. Behav. 98, 277–284. https://doi.org/10.1016/j.chb.2019.04.019.

Shneiderman, B., 2020. Human-centered Artificial Intelligence: Three fresh ideas. AIS Trans. Human-Comput. Int. 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 Quart. 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. J. Assoc. Inform. Syst. (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. Eur. J. Inform. Syst. (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. Inf. Organ. 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. J. Strateg. Inf. Syst. 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. Bus. Horiz. 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. J. Assoc. Inform. Syst. (JAIS) 13, 1–30. https://doi.org/10.17705/1jais.00284.

Whittlestone, J., Nyrup, 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. Harv. Bus. Rev. 96, 114–123.

Winstanley, D., Woodall, J., Heery, E., 1996. Business ethics and human resource management: Themes and issues. Pers. Rev. 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 II

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)*, 1–87.

https://doi.org/10.1007/s10796–025–10596–0

Information Systems Frontiers
https://doi.org/10.1007/s10796-025-10596-0

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

Teresa Hammerschmidt[1] · **Alina Hafner**[2] · **Katharina Stolz**[3] · **Nina Passlack**[4] · **Oliver Posegga**[5] · **Karl-Heinz Gerholz**[4]

Accepted: 2 March 2025
© The Author(s) 2025

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

✉ Teresa Hammerschmidt
teresa.heyder@uni-bamberg.de

Alina Hafner
alina.hafner@tum.de

Katharina Stolz
katharina.stolz@bwi.uni-stuttgart.de

Nina Passlack
nina.passlack@uni-bamberg.de

Oliver Posegga
oliver.posegga@uni-bamberg.de

Karl-Heinz Gerholz
karl-heinz.gerholz@uni-bamberg.de

[1] University of Bamberg, An Der Weberei 5, 96047 Bamberg, Germany

[2] TUM School of Computation, Information and Technology, Technical University of Munich, Arcisstraße 21, 9, 80333 Munich, Germany

[3] University of Stuttgart, Keplerstraße 17, 70174 Stuttgart, Germany

[4] University of Bamberg, Kärntenstraße 7, 96052 Bamberg, Germany

[5] University of Bamberg, An Der Weberei 5, Bamberg, Germany

1 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

Published online: 31 March 2025

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 categorial 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.

2 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 1 (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.

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)

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?	*Groundwork of The Metaphysic of Morals* by Immanuel Kant (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?	*Utilitarianism* by John Stuart Mill (Mill et al., 1989); *An Introduction to the Principles of Morals and Legislation* by Jeremy Bentham (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?	*Nicomachean Ethics* (c. 340 BCE) by Aristotle (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?	*Leviathan* by Thomas Hobbes (Malcolm, 1996); *The Social Contract* by Jean-Jacques Rousseau (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?	*A Theory of Justice* by John Rawls (Rawls, 1971)

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.

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.

3 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.

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 2, Figure 2-4).

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 3, 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 4, 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 5 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 5) 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.

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)

Table 2 (continued)

Concept	Definition	Context: Human Beings	Context: AI Technologies	Intersections with Other Concepts
Moral Patiency (Moral Patients)	Moral patiency 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 patiency intersects with moral and ethical agency, as entities recognized as moral patients are granted moral consideration even if they lack agency (Shevlin, 2021)

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 6, 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 7 (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

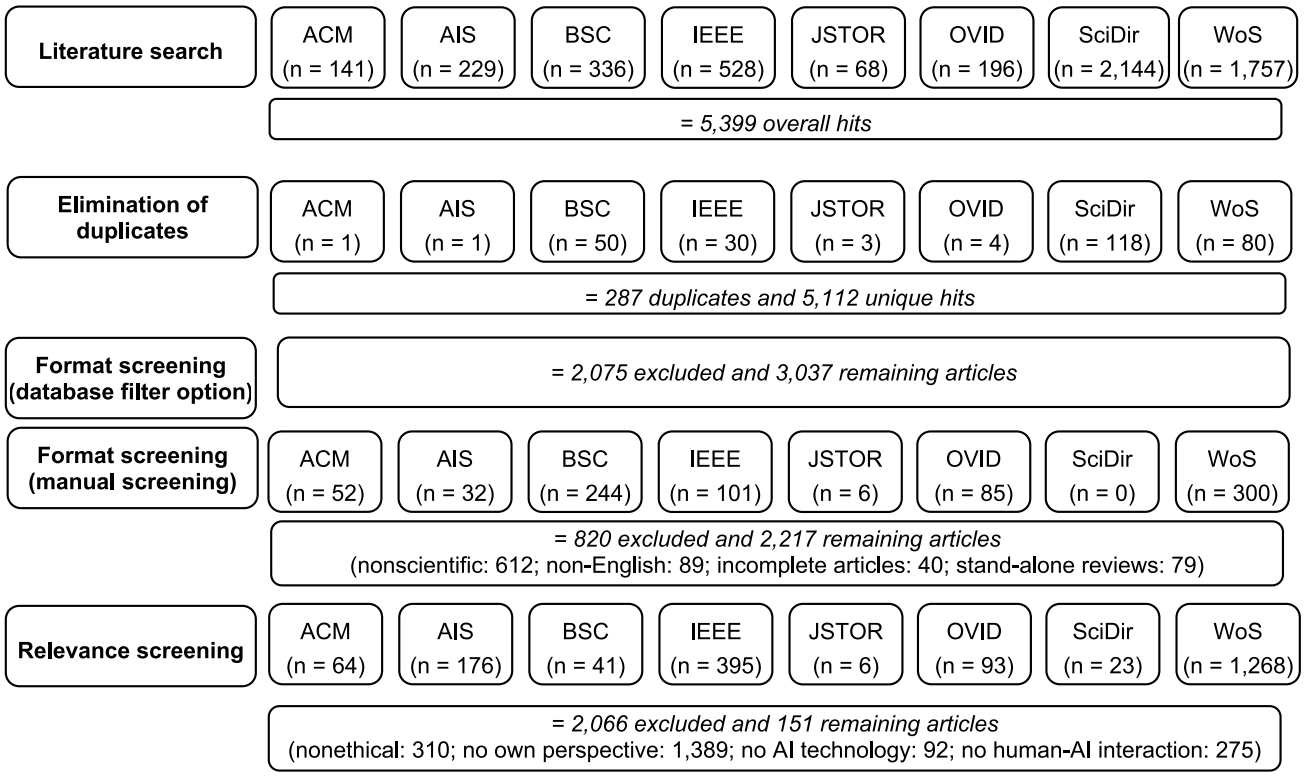


Fig. 1 Search process

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 8 for details).

4 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:

How the AI Ethics Discourse can be Fragmented into Two Main Streams The citation network analysis (see Appendix 9, Figure 5) 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 (Floridi et al., 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

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 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

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 and 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 and Veloso (2022)
All three most common theories (deontology, virtue ethics, and consequentialism/utilitarianism)	16	17.20%	Génova et al. (2023) or Telkamp and Anderson (2022)
Deontology and virtue ethics	19	20.43%	Bankins and 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 Crnkovi & Cürüklü (2012)

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 10, Figures 6-7), 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 11). 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 coauthorships 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 12). 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 (Volkman and Gabriels, 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 13. 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 (Dyrkolbotn et al., 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 and 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 & Scheutz (2018) argue that human control aftersward 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).

5 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., Rismani & 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 Niederman 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 Sanders (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 14.; 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-Bbased 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.

6 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.

Appendix 1. 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, 1992, 2018). *Normative ethics*. Routlegde
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 (1990). *Ethics: Inventing right and wrong*. 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 (2005). *Morality: Its nature and justification.* 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 absen (Anderson & Anderson, 2007)	Singer, P. (Ed.). (1968). *Applied ethics.* Oxford University Press

Appendix 2. Bibliometric Word Analysis

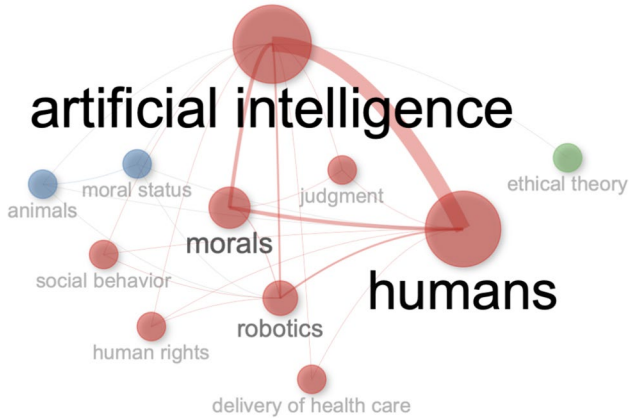


Fig. 2 Most often connected terms (n = 6,523)

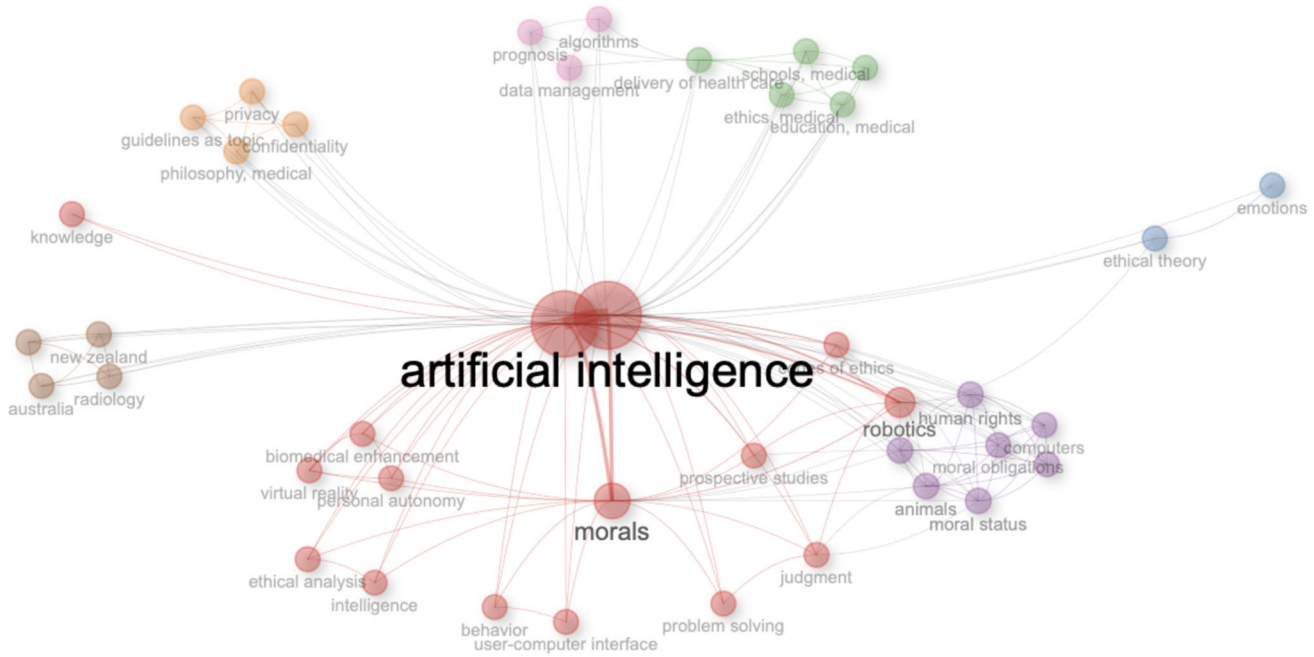


Fig. 3 Word families (n = 6,523)

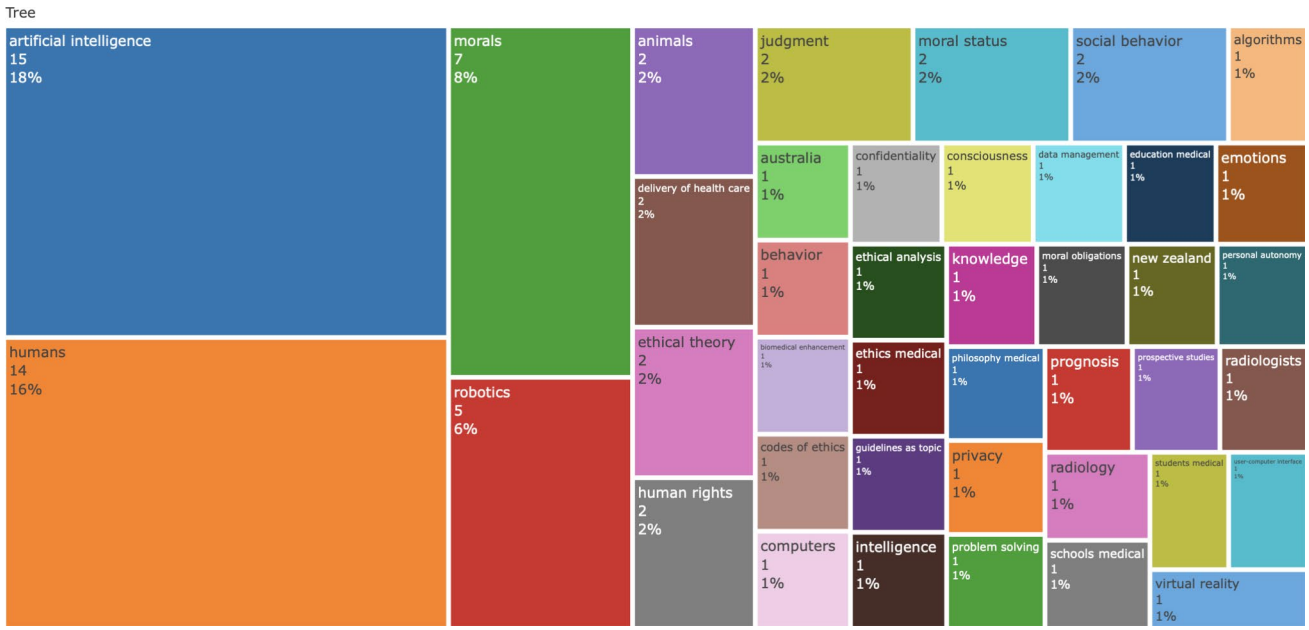


Fig. 4 Keyword appearance (n = 6,523)

Appendix 3. 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 4. 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 alignign 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 5. 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	*AI Responsibility* 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	*Normative Ethics* The work underpins their ethical requirements based on prescribed ethical principles, which can be linked to deontology (implicitly addressed)	*HMA & AMA: hybrid moral agency* 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)	*Qualitative empirical work* 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	*Autonomous robots* 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)	*Normative Ethics* Three ethical views are pointed out in order to approach an artificial moral agency (deontology, consequentialism, and virtue ethics are explicitly addressed)	*HMA & AMA: hybrid moral agency* 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)	*Philosophical essay* This essay discusses artificial moral agency using different perspectives
Arnold and Scheutz (2018) "*The Big Red Button is Too Late: An Alternative Model for Ethical Evaluation of AI Systems*" Ethics Information Technology	*Human as a designer* 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	*Applied ethics* 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	*HMA: shared responsibility* 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	*System design* 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	*Robotic law for AI-AI interaction* The paper discusses the idea of a robotic law, which tackles the point that future AI will interact among themselves	*Normative Ethics* The work makes use of deontological ethics and contractarianism as they further base it on ethical considerations, such as right or wrong	*HMA: full ethical agents* The authors noted that humans have to be in charge of the moral reasoning behind the proposed law	*Position paper* The authors elaborate on their ideas of robotic law
Ashrafian (2015c) "*Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights*" Science Engineering Ethics	*Responsibility of robots* The paper discusses the responsibilities of robots. In the future, robot-robot-interaction will especially be the focus of research	*Meta-ethics* The author discusses the law on ethical AI from a meta-theoretical perspective	*AMA & HMA: hybrid agency* 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 Artificial Moral Agents (AMAs)	*Philosophical essay* The work discusses ethical issues and unclear responsibilities arising with the use of AI for the military

Table 8 (continued)

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
Auld et al. (2022) "*Governing AI through ethical standards: learning from the experiences of other private governance initiatives*" Journal of European Public Policy	*Learning from governance initiatives* This study considers on how to learn from private governance initiatives to guide AI governance	*Applied ethics (implicit)* The paper discusses AI ethics from a regulatory perspective, learning from private governance initiatives	*HMA: full ethical agents and shared responsibilities* The study takes perspective of public–private governance and AI governance and puts responsibility to humans, such as representatives of certain states	*Philosophical essay* The authors make suggestions on how to govern AI based on a discourse from private governance initiatives
Badea and Artus (2022) "*Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents*" LNAI Proceedings	*AI's intentions and morality* This study takes a perspective the interpretation problem on AI and on mistakes which an artificial moral agent could make and categorizes them	*Meta-ethics* 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	*AMA: explicit moral agent* 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)	*Position paper* The authors demonstrate that there is an interpretation problem with AI and AI's moral intention
Bankins and Formosa (2023) "*The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work*" Journal of Business Ethics	*Change of the meaning of work* 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)	*Normative ethics* 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	*No AMA, only HMA* The authors argue that there is no moral agency; human autonomy should be saved (to ensure their well-being)	*Philosophical essay* The paper discusses the role of AI at work
Behdadi and Munthe (2020) "*A Normative Approach to Artificial Moral Agency*" Minds & Machines	*AI entities and responsibility* 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	*Normative ethics* The paper takes a normative lens on the AMA discussion by, in particular, focusing on deontological ethics	*HMA & AMA: hybrid moral agency* 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	*Philosophical essay* The authors illustrate a normative approach abandoning the theoretical discussions on moral agency
Bilal et al. (2021) "*Trust Development in Artificial Intelligence-based Emerging Technologies: Rise of Technomoral Virtues and Data Ethics*" ACIS Proceedings	*Trust in AI* This work studies the most important virtues to develop trust in AI (from individual perspective)	*Normative Ethics* It explicitly refers to virtue ethics theory and uses virtues as concepts	*Combining HMA and AMA* 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	*Mixed-method* This study combines quantitative and qualitative elements using Q-methodology with 39 participants
Bleher and Braun (2023) "*Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice*" Science and Engineering Ethics	*Practical AI ethics* The paper reviews the theory–practice gap in AI ethics and produces three dimensions	*Applied ethics* The paper integrates three reflective dimensions that can help to put AI ethics into practice	*HMA: full ethical agents* The authors argue that as there is no AI moral agency, human moral agency is required for oversights, control, and judgment	*Philosophical essay* This study reflects on AI governance practices

Table 8 (continued)

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
Bogosian (2017) "*Implementation of Moral Uncertainty in Intelligent Machines*" Minds & Machines	*Uncertain machines* 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	*Normative ethics* The paper explicitly considers deontological and virtue ethics	*Building AMA based on HMA:* 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	*Philosophical essay* 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) "*A Case for Machine Ethics in Modeling Human-Level Intelligent Agents*" PHILPAPERS Proceedings	*Human-Level intelligence?* This study discusses the idea of technological singularity and whether greater human-level intelligence will still exist	*Normative ethics* 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	*AMAs as full ethical agents* The authors argue that AI systems are capable of moral reasoning, judgment, and decision-making	*Philosophical essay* The work discusses that different AI researchers use different frameworks for developing AMAs
Bryson (2018) "*Patiency is not a Virtue: The Design of Intelligent Systems and Systems of Ethics*" Ethics and Information Technology	*Societal perspectives* This study discusses recommendations for maintaining either human or artefact-centered societies based on how a society is forming ethics	*Normative ethics* 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	*HMA to build AMA* The authors argue that values generally come from humans' evolutionary origins– humans give the values to AI	*Position paper* 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 and Fieseler (2023) "*Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence*" Business Ethics Quarterly	*Governing AI* 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	*Applied ethics* 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	*HMA: shared responsibility* Responsibilities lie by humans for oversight, control, and so on	*Philosophical essay* 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) "*A sector-based Approach to AI Ethics: Understanding Ethical Issues of AI-related Incidents within their Sectoral Context*" AIES Proceedings	*Stakeholder's perspectives on AI ethics* The study applies a sector-specific perspective trying to understand AI ethics from the stakeholders' view	*Normative ethics* Their research is guided by normative principles. Thus, deontology is implicitly used	*HMA: shared responsibilities* The authors note that there is a form of shared responsibility for different stakeholders in different industries facing similar ethical branches	*Qualitative content analysis* The study 125 incident cases from police, education, politics, automotive, and healthcare
Buruk et al. (2020) "*A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence*" Medicine Health Care and Philosophy	*Developing guidelines* This study analyzes documents with respect to developing ethical norms and guidelines for the planning, development, production, and usage of AI technology in healthcare	*Normative ethics* The work analyzes already produced guidelines and policy documents on AI ethics (contractarianism)	*HMA: shared responsibility* The authors argue that humans need to share their responsibilities to deal with ethical issues surrounding AI and to increase AI's trustworthiness	*Qualitative content analysis* This work is analyzing principles in the healthcare context to determine ethical dilemmas of practical life
Carneiro and Veloso (2022) "*Ethics, Transparency, Fairness and the Responsibility of Artificial Intelligence*" DITTET Proceedings	*Arising ethical problems with AI adoption* 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)	*Normative ethics* The work discusses AI ethics taking a regulatory perspective by implementing AI contracts and policies, which can be linked to contractarianism	*HMA, no AMA* The authors are trying to solve ethical problems from a technical standpoint (referring to robotic laws)	*Syntheses* The paper presents the regulation, methods, and techniques that can help to address ethical problems raising up with the use of AI
Champagne and Tonkens (2023) "*A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot Harm*" Science and Engineering Ethics	*Moral answerability* This study produces a framework for assigning moral answerability for robotic actions to humans and robots into categories. They extend the blank check solution	*Meta-ethics* The paper takes different meta-theoretical perspectives on ethics and discusses the blank-check solutions from these (writing about meta-ethical fictionalism or instrumentalism) *Normative ethics* The paper addresses principles to ensure moral answerability, thereby implicitly taking a deontological perspective on it	*HMA: humans remain responsible* 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	*Conceptualization* 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)

Table 8 (continued)

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
Chaput et al. (2021) "*A Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior*" AIES Proceedings	*Ethical behavior* 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	*Normative ethics* 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 Categorial Imperative)	*Hybrid moral agency* The authors propose a hybrid approach, where AI is guided by some moral frameworks and by adhering to Kant's Categorial Imperative	*System design* This study develops a formal model and experiments on a hypothetical Smart Grid use case in electricity production
Coeckelbergh (2020) "*Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability*" Science and Engineering Ethics	*Responsibility attribution* This study focuses on responsibility attribution in contexts where humans interact with AI agents	*Normative ethics* The authors argue through virtue ethics as they use two Aristotelian conditions for responsibility	*Limited HMA and humans as moral patients* 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	*Philosophical essay* The work discusses responsibility questions of AI actions in the context of moral agency and moral patience
Colmenarejo et al. (2022) "*Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation*" AIES Proceedings	*AI governance through different perspectives* This study investigates AI regulation from an interdisciplinary perspective by considering how law and industries achieve agreement with European values	*Normative ethics* The study analyzes AI regulation from different perspectives that can be linked to various ethical views, such as deontological ethics and consequentialism	*AMA should be regulated and based on HMA* The authors argue that European values can guide AI design and extend morality within AI systems	*Position paper* 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) "*Moral Decision Making Frameworks for Artificial Intelligence*" AAAI Proceedings	*Automating moral judgment* 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)	*Normative ethics* 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	*HMA: shared responsibility* 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	*Philosophical essay* 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) "*Understanding Responsibility in Responsible AI. Diagnostic Virtues and the Hard Problem of Context*" Ethics and Information Technology	*Developing moral AI* 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)	*Normative ethics* The study's framework is based on Aristotelian virtue, as it plays a big role in terms of AI responsibility and human responsibility	*HMA: full ethical agents* 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) *AMA: implicit ethical agents* The authors argue that AI technologies can demonstrate some forms of diagnostic virtue following a rule-based form of morality given through humans	*Conceptualization* 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) "*Robust Artificial Moral Agents and Metanormativity*" AIES Proceedings	*Metanormativity is required for the design of ethical AI* 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	*Meta-ethics* 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)	*AMAs: full ethical agents* The authors argue that AI is capable of ethical reasoning	*Philosophical essay* The paper explores the relationship between our ignorance concerning certain meta-normative topics and the design of ethical artificial intelligence (AI)

Table 8 (continued)

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
Cooper et al. (2022) "*Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning*" FACCT Proceedings	*Responsibility attribution* 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	*Meta-ethics* The work grounds on moral philosophy by using Nisenbaum's' barriers in moral philosophy	*HMA: full ethical agents* The authors argue that humans are accountable for actions (e.g., for bugs) and, thus, also for the ethics put into the machines	*Philosophical essay* The paper discusses Nisenbaum'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) "*Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens*" ACM HCI Proceedings	*Emotional AI* 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)	*Normative ethics* 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)	*Limited HMA* 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	*Quantitative online survey* The study examines the perceived benefits and risks of 395 US adults associated with emotional AI use at the workplace
Córdova and Vicari (2022) "*Practical Ethical Issues for Artificial Intelligence in Education*" TECH-EDU Proceedings	*Designing AI curricula* 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	*Normative ethics* 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	*AMA can be used to strengthen HMA* 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 contributeing to HMA)	*Conceptualization* 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 Crnkovi & Cürüklü (2012) "*Robots: Ethical by Design*" Ethics Information Technology	*Creating artificial morality for ethical behavior* This study claims that artificial morality should be built into future robots/softbots, with the aim of ensuring their ethically adequate behavior	*Normative ethics* The paper investigates how ethical robots should be designed by principles and policies (deontology and contractarianism)	*HMA: full ethical agents* The authors argue that humans are in charge of 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)	*Philosophical essay* This work discusses guidelines for robots' ethics by design
Dyrkolbotn et al. (2018) "*On the Distinction between Implicit and Explicit Ethical Agency*" AIES Proceedings	*Implicit and explicit agents* 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)	*Normative ethics* The work claims that moral values are explicitly given through existing legal categories that provide boundaries by law, thus implicitly addressing contractarianism	*HMA: shared responsibility* 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)	*Philosophical essay* 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

Table 8 (continued)

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
Eicher et al. (2018) "*Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies*" AIES Proceedings	*Ethical issues of virtual assistants* 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	*Descriptive ethics* 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	*HMA as ethical baseline* 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	*Case study* 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) "Towards A Unified Utilitarian Ethics Framework for Healthcare Artificial Intelligence" AMCIS Proceedings	*AI guidelines for ethical AI* 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)	*Normative Ethics* They adopt a utilitarian framework in designing ethical AI in healthcare (consequentialism)	*Build AMA based on HMA* The authors argue that ethical AI can be designed based on ethics frameworks made, discussed, and evaluated by humans	*Interviews* The study conducts a secondary data analysis based on interviews and survey data with AI experts
Erman and Furendal (2022) "*The Global Governance of Artificial Intelligence: Some Normative Concerns*" Moral Philosophy and Politics	*Societal impact of AI striving for global AI governance* 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	*Normative Ethics* 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	*HMA is broader than AMA* 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	*Position paper* The study discusses global AI governance norms and arising normative concerns
Estrada (2018) "*Value Alignment, Fair Play, and the Rights of Service Robots*" AIS Proceedings	*Moral behavior* 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)	*Meta-ethics* The paper takes a metatheoretical perspective on the concept of the Moral Turing Test	*Influence of AMA on HMA* The authors argue that AI's moral behavior can impact humans because Moral Turing Tests imitate human moral behavior	*Conceptualization* 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 and Etzioni (2017) "*Incorporating Ethics into Artificial Intelligence*" Journal of Business Ethics	*Ethical choices in AI design* 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	*Meta-ethics* 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	*Integrating HMA in AI is better than developing AMA* 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	*Philosophical essay* 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	*Data protection* 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)	*Normative ethics* 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) *Applied ethics* The work deals with the transition from written regulations such as laws to practice	*HMA: shared responsibility of humans* 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	*Philosophical essay* This work makes propositions for future work and deals with the relation between AI, data protection, ethics, intelligent systems, privacy, and robotics

Table 8 (continued)

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
Farisco et al. (2020) "*Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence*" Science and Engineering Ethics	*Ethical analysis requires theoretical and practical considerations* 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	*Meta-ethics* 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	*HMA: full ethical agent* 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	*Philosophical essay* 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	*Issues in achieving responsible AI practices* 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)	*Applied ethics* The study argues that ethical frameworks need to be transferred into an organizational culture to be effective *Normative ethics* The study evaluates and analyses several ethical principles including implicitly addressing consequentialism and deontology	*Aligning AMA with HMA* The authors argue that to overcome responsibility issues in practice, AI must be designed to be trustworthy and aligned with key ethical guidelines	*Interview* 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) "*Imagine a More Ethical AI: Using Stories to Develop Teens' Awareness and Understanding of Artificial Intelligence and its Societal Impacts*" IEEE Proceedings	*Awareness of AI biases* 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	*Descriptive ethics* The study carries out an experiment with students to determine their perceptions and understanding of (ethical) AI	*Strengthening HMA* 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	*Experiment* This work performs a classroom experiment in the context of informal and formal learning settings
Fukuda-Parr & Gibbson (2021) "*Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines*" Global Policy	*Social good vs. transparency: weighting ethical principles* This study discusses that most ethical frameworks focus on transparency instead of principles to ensure the social good, such as discrimination, inequalities, and participation	*Normative ethics* 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	*HMA: shared responsibility* The authors argue that since governments and civil society are responsible for developing AI standards, ethical principles are rooted in international human rights	*Syntheses* The work analyzes 15 guidelines on ethical practices, whereby the definition of inequality and ethics should be widened
Gahnberg (2021) "*What Rules? Framing the Governance of Artificial Agency*" Policy and Society	*AI governance framing* 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	*Normative ethics*: 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)	*AMA: implicit ethical agents* The author argues that AMA would involve the ability to make decisions and exhibit behavior that are opaque and unpredictable, which requires specific regulation	*Philosophical essay* 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) "*Machine Ethics: Do Androids Dream of Being Good People?*" Science and Engineering Ethics	*Machine ethics* This work sees machine ethics as part of moral philosophy, which deals with ensuring the ethical behavior of machines with AI	*Normative ethics* This work implicitly presents ideas that can be linked to virtue, deontological, and utilitarian ethics	*AMA: implicit ethical agents* 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	*Position paper* 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

Table 8 (continued)

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
Gerdes (2022) "*A Participatory Data-Centric Approach to AI Ethics by Design*" Applied Artificial Intelligence	*Focus on data when designing ethical AI* 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	*Normative ethics*: 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	*HMA: shared responsibility* 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	*Conceptualization* This work addresses ethical challenges when designing and developing systems with a focus on developers and experts working together
Ghatar et al. (2023) "*Practices for Responsible AI: Findings from Interviews with Experts*" AMCIS Proceedings	*Responsible AI practices* This work identifies practices for explainability, fairness, and inclusiveness of AI systems	*Applied ethics* The work investigates the extent to which the principles for responsible development and deployment of AI have been operationalized	*HMA building/training AMA*: The authors argue that humans are in charge of integrating responsible AI practices to ensure, for example, fairness and morality	*Interviews* 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) "*Artificial Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI*" Journal of Business Ethics	*Machine ethics and morality* This work describes machine morality as dealing with the possession and implementation of morality and, especially, the moral capacity of autonomous machines	*Descriptive ethics* This study conducts a meta-analysis of three previous experiments to investigate the interaction with and moral concerns of humans towards technologies vs. humans	*HMA-AMA: distributed moral responsibility* 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	*Experiment* This study contains a meta-analysis of 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	*Social justice with ethical AI* 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	*Applied ethics* The work refers to several public partnerships, ethical boards, and declarations for responsible AI	*HMA: shared responsibility of specific groups of people* 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	*Conceptualization* 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

Table 8 (continued)

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
Gómez de Ágreda (2020) "*Ethics of Autonomous Weapons Systems and its Applicability to any AI Systems*" Telecommunications Policy	*Control over AI* 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	*Normative ethics* 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	*HMA: full ethical agents* 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)	*Syntheses* 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	*Legal implementation of ethical codes* 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	*Descriptive ethics* The work uses interviews to identify public values and sociocultural norms	*HMA: full ethical agents* The author highlights that human morals and norms serve as a base for ethical guidelines	*Case study* 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) "*The Ethics of AI Ethics: An Evaluation of Guidelines*" Minds & Machines	*AI ethics* 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	*Normative ethics* 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	*HMA: full ethical agents* The author highlights that every individual should generate motivation and adopt habits to use AI in a responsible and positive manner	*Syntheses* 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) "*Machine Ethics, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions?*" Society	*Pluralism of AI ethics* 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	*Normative ethics* 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 *Applied ethics* 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	*AMA: implicit ethical agents that depend on HMA* The author introduces the concept of "polymorphic recursive allostery" 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	*Position paper* 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

Table 8 (continued)

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
Hauer (2022) "*Incompleteness of moral choice and evolution towards fully autonomous AI*" Humanities & Social Sciences Communications	*Ethics of fully autonomous AI* 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	*Normative ethics* 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	*Human moral agency* The author argues that AI ethics covers several sub-questions, including the establishment of institutions that implement moral issues in research	*Philosophical essay* The authors discuss AI's full automation of tasks, including the transfer of moral competence
Hawkins and Mittelstadt (2023) "*The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices*" FACCT Proceedings	*The gap between norms and practices* 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	*Normative ethics* The paper presents research ethics policies that can be linked to contractarianism *Applied ethics* The study investigates how data enrichment can close gaps in the context of crowdsourcing	*HMA: shared responsibility* The study highlights gaps between ethics norms and practices, encouraging shared responsibilities between users and regulators	*Quantitative text analysis* The work deals with a gap regarding ethical disclosures in research and ethical issues while working with crowdsourced workers
Henderson (2019) "*Teaching Data Ethics*" CEP Proceedings	*AI curricula* This study aims to make students aware of 1) algorithmic accountability, 2) bias in AI, 3) data protection and filter bubbles, and 4) practical mechanisms for dealing with this topic	*Applied ethics* The work presents a new module on Data Ethics and Privacy (AI curricula)	*HMA: training human moral agency* The authors argue that humans should be trained to be aware of ethical issues in the AI context	*Case study* The study presents a new "Data Ethics & Privacy module" for students of computer science in 2018 in the UK
Hooker and Kim (2018) "*Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic*" AIES Proceedings	*Logic vs. intuition of AI ethics* This work aims to achieve breakthroughs in issues that cannot be resolved by human ethicists	*Normative ethics* 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	*HMA-AMA: distributed moral responsibility* 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	*Conceptualization* 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) "*AI in Interpreting: Ethical Considerations*" Across Languages and Cultures	*HCAI as a moral obligation* 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	*Normative ethics* The study mentions the European General Data Protection Regulation (GDPR). Thus, contractarianism is implicitly addressed *Meta-ethics* The study provides general thoughts on AI, AI ethics, and ethical principles (data bias and data quality, data privacy and data ownership, transparency)	*Enhance HMA with AMA* 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	*Position paper* 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) "*Building a Maturity Model for Developing Ethically Aligned AI Systems*" IRIS Proceedings	*Guidelines for ethics communication* The study highlights that guidelines are often used to communicate ethical intentions, whereby no single standard has emerged	*Normative ethics*: The study argues that AI ethics guidelines should be used as a base when developing and designing ethical AI systems (deontology; implicitly addressed)	*HMA as baseline* The authors argue that AI ethics guidelines serve as a base for ethical AI systems	*Syntheses* 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

Table 8 (continued)

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
Jobin et al. (2019) "*The Global Landscape of AI Ethics Guidelines*" Nature Machine Intelligence	*Landscape of ethical principles* 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	*Normative ethics* The study introduces and analyses numerous guidelines as a base for ethical AI, which can be linked to deontology *Applied ethics* The study structures existing ethical AI guidelines	*HMA as baseline* The authors argue that AI ethical guidelines created by humans are the base for practitioners to achieve ethical AI applications	*Syntheses* The work structures and organizes ethical guidelines, principles, and norms of several institutions
Katznelson and Gerke (2021) "*The Need for Health AI Ethics in Medical School Education*" Advances in Health Science Education	*Educate AI ethics for healthcare* 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	*Applied, normative, and meta-ethics* 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	*Training HM*: The authors argue that medical students should be trained to develop, implement, engage with, and evaluate AI in a meaningful and ethical way	*Conceptualization* 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 Ontology	*From AI ethics to medical ethics* 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	*Normative ethics* The study analyzes ethics and standards in the context of using AI in medicine. Thus, deontology is implicitly addressed	*Training HMA* The authors argue that standards and principles alone are not enough, as people need to be upskilled	*Synthesis* 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	*Applying AI ethics* 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	*Applied ethics* In the study, several AI practitioners and lawmakers were invited to complete an online questionnaire	*HMA as baseline* The authors highlight that standards are solely developed by human beings, providing the baseline for ethical conduct	*Quantitative online survey* 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	*Perceptions of ethical AI* 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)	*Normative ethics* The study focuses on ethical design principles, which can be linked to deontology *Descriptive ethics*: The study contains a survey of 1,204 people to gain insights into their opinions and perceptions of AI ethical theory	*Using HMA to build AMA*: The authors argue that ethical principles are needed to mathematically integrate ethics into the specific algorithm and AI application	*Quantitative online survey* 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	*Moral decision-making models* The study discusses a model for moral decision-making in the context of autonomous vehicles	*Descriptive ethics* The study uses the hierarchical Bayesian model, whereby moral values are analyzed empirically	*HMA-AMA: Distributed moral responsibility*: 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	*Conceptualization* The authors present a computer model to develop the moral judgment capability of autonomous vehicles based on empirical insights

Table 8 (continued)

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
Knox (2023) "*(Re)politicising data-driven education: from ethical principles to radical participation*" Learning, Media and Technology	*Ethics is not solely a technical matter* 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	*Normative ethics* 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)	*HMA: shared responsibility* 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	*Position paper* 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	*Influence of AI agents on human's ethical behavior* 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)	*Normative ethics* 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	*AMA impacts HMA* 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	*Conceptualization* 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	*Shift from ethics to governance* 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	*Normative ethics* 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	*HMA: moral responsibility* 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	*Syntheses* 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	*Kant in AI curricula* 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	*Applied ethics* The work deals with Kant's pedagogy	*HMA defines AMA* 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	*Philosophical essay* 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	*Ethics vs. morality* 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)	*Applied ethics* The study discusses several principles for the use of service robots in classrooms	*Strengthening HMA* 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	*Syntheses* 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	*Dilemmas of managerial AI* 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	*Normative ethics* This work introduces the deontological school of managerial responsibility for AI, including explicit rules and principles	*Replacing HMA through AMA* 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	*Philosophical essay* The study describes philosophical ideas on the ethical dilemma regarding the application and promotion of AI is addressed

Table 8 (continued)

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
Lara (2021) "*Why a Virtual Assistant for Moral Enhancement When We Could have a Socrates?*" Science and Engineering Ethics	*Machine ethics* 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	*Applied ethics* The study concerns whether technological advances should be used to optimize and improve people as moral agents *Normative ethics*: 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)	*Train HMA with AI* 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	*Conceptualization* 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	*AI mimicking friendship* 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 human-like friendly AI, mimicking friendship	*Applied ethics* The study discusses how AI should be designed and developed as a friendly AI, including, for example, mimic aspects	*AI impact HMA* 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	*Philosophical essay* The work deals with friendly AI, which includes mimicking issues of human friendship
Lim and Kwon (2021) "*A Study on the Modeling of Major Factors for the Principles of AI Ethics*" DG.O Proceedings	*Standardization of AI principles* This study investigates major factors of the different ethical principles of AI ethics that exist across several countries to identify some standards	*Normative ethics*: The study presents principles on what is morally, thus addressing deontology implicitly *Applied ethics* The study analyzes the ethical principles of 15 different countries	*HMA: shared responsibilities* The authors illustrate that AI ethic principles are presented by different governmental institutions	*Conceptualization* 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	*Ethical issues of human-like AI* This work provides moral considerations on ethical issues that arise when AI increasingly performs human-like cognitive processes	*Normative ethics* 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	*HMA guiding AMA* The author argues that ethical AI should be determined by human values as AI lacks consciousness	*Position paper* 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	*Robots as moral agents* This work is concerned with the development of robots and other systems that are capable of making ethical and moral decisions	*Meta-ethics* This work describes computational and philosophical facets of artificial moral agency	*Rational agency* 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)	*Philosophical essay* The work provides a combination of a philosophical and scientific perspective on artificial moral agency

Table 8 (continued)

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
Manna and Nath (2021) "*Kantian Moral Agency and the Ethics of Artificial Intelligence*" Problemos	*Kantian moral agency* 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)	*Meta-ethics* The work discusses whether AI agents can have a Kantian moral agency	*AI simulates HMA* 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	*Philosophical essay* 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	*Epistemic AI ethics* The work addresses concerns that decisions are made on the basis of incorrect or inconclusive findings	*Normative ethics* 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)	*HMA: shared responsibilities* The authors argue that human-centric AI has been the focus, leading to several AI ethics principles	*Qualitative content analysis* 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*" Journal of Business Ethics	*Communicating ethical implications* Information on algorithm accountabilities during development and implementation processes is crucial	*Normative ethics* The work implicitly addresses consequentialism by stating, "In that algorithm create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity" (p. 835)	*AMA-HMA: Distributed moral responsibility* The authors criticized that algorithm's structure impacts the lives of human beings, whereby the responsibility is not finally defined	*Position paper* This work focuses on the opinion that algorithmic has a huge impact on human beings, 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	*Machine ethics* 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	*Descriptive ethics* The study investigates moral uncertainties empirically	*Rising AMA*: The authors argue that AI systems are increasingly acting autonomously, leading to greater (moral) agency	*Quantitative online survey* 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	*Human-centered AI implementation* The study implies that the user plays a significant role in fostering morality in human-AI interaction	*Applied ethics* The study focuses on how companies can encourage the implementation of AI ethics *Normative ethics* 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)	*HMA guides AI* The authors argue that human values should guide AI's self-learning and autonomous operations	*Interviews* The study uses nine semi-structured expert interviews to identify measures and mechanisms to encourage employees for ethical AI practices

Table 8 (continued)

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
McDonald and Pan (2020) "*Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact*" PACM Proceedings	*Student's perceptions on AI and ethics* 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)	*Normative ethics* 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))	*HMA-AMA: shared/distributed moral responsibility* 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)	*Interviews* 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	*Hybrid ethics* 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	*Applied ethics* 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)	*Distributed moral responsibility: explicit ethical agents* 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)	*Synthesis* The main objective of the paper is a design tool aimed at preventing the development of robots that impose morally questionable paternalism on users
Minkkinen and Mäntymäki (2023) "*The Institutional Logics Underpinning Organizational AI Governance Practices*" SCIS Proceedings	*AI governance* 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	*Normative ethics* 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	*AMA: shared responsibility* The paper focuses on how to translate AI governance and ethical principles to AI so that AI acts according to human values	*Synthesis* 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) "*Ethics as a Service: A Pragmatic Operationalisation of AI Ethic*" Minds & Machines	*Operationalizing ethics* 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)	*Applied ethics* The study aims to make AI ethics useful for practitioners by referring to a concept that the authors call "ethics as a service."	*Hybrid moral agency:* 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)	*Synthesis* 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) "*The Ethical Implications of Using Artificial Intelligence in Auditing*" Journal of Business Ethics	*Humans in control* 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)	*Applied ethics* The paper focuses on ethical issues arising in practice, while the paper relies on ethical principles and codes of conduct	*Hybrid moral agency: implicit ethical agents* 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)	*Synthesis & case study* 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

Table 8 (continued)

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
Neubert and Montañez (2020) "*Virtue as a Framework for the Design and use of Artificial Intelligence*" Business Horizon	*Relevance of virtue ethics* This study highlights that virtue ethics performed by humans is needed to guarantee ethical human-AI interaction	*Applied ethics* The study highlights real-world scenarios in which they transfer their views on virtue ethics *Normative ethics* 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)	*HMA: humans remain responsible* 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	*Quantitative text analysis* 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) "*Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers!*" KES Proceedings	*Mixing ethical views* 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	*Applied ethics* The study outlines ethical guidelines for AI developers that are designed according to ethical principles *Normative ethics* The study compares and combines virtue ethics and deontology	*AMA: explicit ethical agents* 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)	*Synthesis* 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 and Baker (2023) "*Ethics and AI Issues: Old Container with New Wine?*" Information Systems Frontier	*Using the Turing test for the development of ethical AI* 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)	*Normative ethics* 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 be linked to different ethical views (p. 10)	*Hybrid moral agency: implicit ethical agents* 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 actions or neglectful inattention rather than explicit, purposeful agency" p. 10)	*Synthesis* 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
Obi and Gray (2023) "*Auditing Practitioner Judgment for Algorithmic Fairness Implications*" IEEE Proceedings	*Judging algorithmic fairness* The paper emphasizes the key role that designers play in their judgment decisions when designing AI	*Descriptive ethics* The study reflects on the role of practitioners in developing ethical AI systems and thereby connects practitioner judgments to related fairness and justice implications	*HMA: shared responsibility* The paper shows the key role that designers play in developing AI	*Case Study* 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	*Different views on ethics* The paper suggests that different stakeholder's views and regulations are relevant for ethical human-AI interaction	*Normative ethics* 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	*HMA: shared responsibility* 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)	*Synthesis* The authors suggest a framework that highlights the implications and consequences of affective-aware AI based on prior research

Table 8 (continued)

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
Palmer and Schwan (2024) "*More Process, Less Principles: The Ethics of Deploying AI and Robotics in Medicine*" Cambridge Quarterly of Healthcare Ethics	*Process over principles* The study highlights that ethics ought to emphasize process over principles; policy outcomes should be influenced by diverse stakeholders	*Applied ethics* 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)	*HMA: moral intention* The paper focuses on the ethical employment of AI as a moral agency rather than with humans	*Position Paper* 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	*Need for AI governance* 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)	*Normative ethics* The study's main focus lies on principles for establishing a moral framework, which is linked to deontology	*AMA: responsibility attribution* 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)	*Synthesis* 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	*Human perceptions on responsibilities* The paper investigates humans' perceptions of responsibility for unethical outcomes	*Descriptive ethics* The paper states that assigning moral agency to AI and taking responsibility depends on individual factors and personality	*HMA-AMA: moral intention* The authors address the fact that humans assign moral agency to AI when making moral judgements	*Experiment* 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 and Zamani (2022) "*Towards Europe's AI Strategy: Desires & Realities*" ECIS Proceedings	*Human perceptions of AI strategies* This work refers to individuals' perceptions of Europe's AI strategy	*Descriptive ethics* The study investigates different stakeholder perceptions of AI frameworks and strategies	*Protecting HMA* The authors perceive AI algorithms as black boxes, suggesting that humans need to keep their autonomy in human-AI interaction	*Qualitative content analysis* 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	*Process over principles* 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	*Applied ethics* 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)	*AMA: explicit agents that require human oversight* 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	*Synthesis* 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	*Interdisciplinary AI curricula* 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)	*Meta-ethics* 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)	*HMA: shared responsibility* The study highlights that "there is an underlying tension between individual agency and bureaucratic control with actual responsibility lying somewhere in between" (p. 517)	*Position Paper* 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

Table 8 (continued)

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
Rismani and Moon (2023) "*What Does it Mean to be a Responsible AI Practitioner: An Ontology of Roles and Skills*" AIES Proceedings	*Skills and roles for responsibilities* 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	*Normative ethics* The study suggests skills, qualifications, etc., that are relevant to ethical practitioners' sense of responsibility, which can be linked to virtue ethics	*HMA: shared responsibilities* The authors focus on human moral agency by suggesting a competency framework splitting responsibilities among AI practitioners based on roles and competencies	*Interviews* 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 and Montoya (2022) "*Contextualizing Artificially Intelligent Morality: A Meta-Ethnography of Theoretical, Political and Applied Ethics*" FICC Proceedings	*Different perspectives on AI ethics require hybrid approaches* 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)	*Meta-ethics* The study follows holistic considerations of AI morality (top down and bottom up) from a theoretical, technical, and political stance	*HMA: no ethical agents* 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)	*Philosophical essay* 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 European Union: Comparing aims and Promoting Ethical Outcomes* " The Information Society	*AI governance: comparing aims with moral values* This work argues that human AI ethics depends on context, culture, and society and should be seen as ethical pluralism guided by policy makers	*Normative ethics* The study argues that moral norms come from mutual agreements (linked to contractarianism), and their interpretation can 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)	*HMA: shared responsibility* The authors suggest to "establish an AI accountability mechanism to clarify the responsibilities of developers, users, beneficiaries, etc." (p. 87)	*Synthesis* By comparing China and the EU, the paper outlines policy suggestions for policymakers to promote ethical 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	*Moral reasoning in HRM* This work focuses on human reasoning in ethical decision-making in HRM processes	*Applied ethics* The study provides ethical considerations in decision-making in HRM *Normative ethics* The study combines different normative views, such as ethical egoism, deontology, utilitarianism, relativism, virtue ethics, and ethics of care, in the HRM context	*AMA: implicit ethical agents* 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)	*Position paper* The paper suggests a model with six dominant algorithmic pathways to provide possible solutions to foster ethical decision-making in HRM
Saariluoma and Leikas (2020) "*Designing Ethical AI in the Shadow of Hume's Guillotine*" IHSI Proceedings	*Ethical AI* 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)	*Applied ethics & normative ethics* The paper raises questions about whether morality can be integrated into AI by reflecting on different ethical views	*AMA: explicit ethical agents* 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)	*Philosophical essay* The paper is questioning to which extent AI can be assigned moral values

Table 8 (continued)

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
Schlimbach and Khosrawi-Rad (2022) "*Towards Ethical Design Features for Pedagogical Conversational Agents*" AMCIS Proceedings	*Pedagogical conversational agents* This study aims to develop requirements and specific features for pedagogical conversational agents that uphold ethical standards	*Applied ethics* The study is concerned with the ethical design of pedagogical conversational agents	*AMA: explicit ethical agents* AI morality is influenced by data and humans, and AI is limited in its agency. Humans are accountable for AI output	*Design science approach* 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	*Individuals' literacy and morality shape AI design* 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)	*Normative ethics* 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	*HMA: moral intention* 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	*Design science research & case studies* 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 and Thorson (2022) "*Computation, Rule Following, and Ethics in AIs*" HICSS Proceedings	*Aligning rules with context* This study focuses on AI alignment and suggests that AI 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)	*Descriptive ethics* The paper introduces rules, including moral and ethics rules, as linguistic entities and illustrates how they can be encoded and treated computationally	*HMA: limited human agency* The authors argue that human ethical and moral reasoning could be understood as fundamentally computational and does not involve some special human-only force or quality	*Philosophical essay* The paper proposes a framework 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	*Implementing ethical AI principles* 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)	*Applied ethics* The study focuses on the implementation of ethical principles in practice	*HMA: moral intention* 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	*Interviews* 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	*AI ethics and machine ethics* The paper examinates 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	*Normative ethics* 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)	*HMA: full ethical agents* 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	*Philosophical essay* 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	*AI perceptions shape ethics in human-AI interaction* 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	*Normative ethic* 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	*HMA: hybrid agency* 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	*Philosophical essay* The work is an essay that strives for the development of realistic AI agents

Table 8 (continued)

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
Shank et al. (2019) "*When are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions after Individual and Joint Decisions*" Information, Communication& Society	*Moral attribution of responsibility within joint decision-making* 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	*Normative ethics* The work implicitly addresses consequentialism, focusing on the ethical implications of AI-driven outcomes and the assignment of responsibility for moral transgressions *Descriptive ethics* The work contains an empirical investigation of responsibility attribution and moral judgment	*HMA to AMA: hybrid agency* 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	*Experiment* 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 Patient?*" Cambridge Quarterly of Healthcare Ethics	*AI robots as moral patients* This work addresses the critical question of when and under what conditions artificial beings might be granted moral status, engaging deeply with the moral patient versus moral agent debate	*Meta-ethics* This work takes an ontological approach, exploring five strategies to determine whether AI can be considered artificial psychological moral patients, focusing on cognitive equivalence with humans	*AMA: hybrid agency* The authors argue that to consider AI robots as moral patients requires attributes such as intuition, empathy, autonomy, and behavioral equivalence. Hence, moral agency depends on the extent humans can draw equivalences to other beings they interact with	*Philosophical essay* The study discusses conditions under which AI-based robots might be considered moral patients in the context of healthcare
Shih et al. (2021) "*Learning Ethics in AI—Teaching Non-Engineering Undergraduates through Situated Learning*" Sustainability	*AI ethics in non-technological classes* 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	*Applied ethics* The study advocates for a situation-based, hands-on learning approach with AI to heighten awareness of AI ethics among students	*HMA-AMA: moral intentionality* 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	*Case study* The paper provides a case study on educational AI initiatives in Taiwan with 328 first-year students with a non-technical background
Shneiderman (2020) "*Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems*" ACM Transactions on Interactive Intelligent Systems	*HCAI guidelines* 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)	*Applied ethics* 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 *Normative ethics* The study implicitly refers to contractarianism as industrial and societal standards of AI governance policies are addressed, which can be further related to deontology	*HMA: distributed agency* 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)	*Conceptualization* 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) "*Towards a Feminist Metaethics of AI*" AIES Proceedings	*AI ethics through a feminist meta-ethical lens* 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	*Meta-ethics* The paper employs a feminist meta-ethical perspective to interrogate the relational aspects of AI ethics, drawing on insights from philosophers with practical experience	*HMA: moral intentionality* 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	*Philosophical essay* The work considers AI ethics from a societal and feministic-European perspective

Table 8 (continued)

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
Siau and Wang (2020) "*Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI*" Journal of Database Management	*AI ethics and ethical AI* 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	*Normative ethics* The paper examines virtue ethics (think critically about moral values) and deontology (ethics as a set of principles) *Applied ethics* The paper further refers to computer ethics and machine ethics as an applied ethics branch	*HMA-AMA: hybrid agency* 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)	*Conceptualization* The paper provides a framework based on the review of AI ethics objectives and ethical principles of eight institutions
Smith and Vickers (2021) "*Statistically Responsible Artificial Intelligences*" Ethics and Information Technology	*Responsible AI* 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	*Normative ethics* The study implicitly engages with virtue ethics through the application of Strawson's (2008) approach to moral responsibility, rooted in human existence and interaction	*AMA: full ethical agents* 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	*Conceptualization* 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	*AI governance: The European AI Act* 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	*Applied ethics* 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 *Normative ethics* The study further implicitly addresses contractarianism referring to AI governance policies	*HMA: distributed agency* 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	*Interviews* 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	*Limitation of AI ethics principles* 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	*Normative ethics* The work refers to all well-known ethical theories of normative ethics.	*AMA-HMA: full ethical agents* 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	*Delphi study* 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

Table 8 (continued)

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
Sullivan and Fosso Wamba (2022) "*Moral Judgments in the Age of Artificial Intelligence*" Journal of Business Ethics	*Responsibility attribution* The study outlines that moral and ethical values are complex and can very between the several interactions between AI and humans. Hence, moral responsibility attribution depends on the relationship between perceived harm and blame judgment	*Normative ethics* The study implicitly refers to virtue ethics as moral responsibility attribution depending on one's inherent moral perceptions and maxims	*HMA-AMA: hybrid agency* 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	*Experiment* 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 and Ye (2023) "*Moral Considerations of Artificial Intelligence*" Science & Education	*Claim for environmental ethics* 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)	*Meta-ethics* The paper refers to environmental ethics as an inclusive form of ethics (explicitly addressed)	*HMA: moral responsibility* 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 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	*Conceptualization* The paper presents a framework for realizing moral AI in education
Susser (2019) "*Invisible Influence*" AIES Proceedings	*AI perceptions* 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	*Meta-ethics* The study explicitly addresses the ethics of invisibility	*HMA: limited agency* 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	*Philosophical essay* This study is a philosophical work framing the choice architecture of AI to avoid manipulation and to increase users' trust
Swanepoel (2021) "*The Possibility of Deliberate Norm-Adherence in AI*" Ethics and Information Technology	*Norm-Adherence* 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)	*Normative ethics* 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)	*AMA: implicit and explicit ethical agents* Swanepoel builds on Moor's (2006) 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	*Philosophical essay* 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

Table 8 (continued)

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
Taeihagh (2021) "*Governance of Artificial Intelligence*" Policy and Society	*Risk management* The paper addresses the governance of AI, particularly in managing risks associated with unexpected and unintended consequences of AI technologies	*Normative ethics* The study implicitly engages with deontological ethics by emphasizing the importance of governance mechanisms that seek to mitigate potential risks associated with AI automation	*HMA: distributed agency* 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	*Position paper* 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) "*AI Ethics and the Banality of Evil*" Ethics and Information Technology	*The banality of evil* 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	*Normative ethics* The work covers various normative ethical theories, including deontology, consequentialism, contractarianism, and virtue ethics	*AMA: implicit ethical agents* 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 principles from a single theory, cannot make genuinely ethical decisions	*Philosophical essay* 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 acts, as they lack the capacity for true moral judgment (p. 447)
Tasioulas (2018) "*First Steps Towards an Ethics of Robots and Artificial Intelligence*" Journal of Practical Ethics	*AI perceptions on ethics of robots* 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	*Normative ethics* The study implicitly engages with deontological ethics through its focus on regulatory standards and with virtue ethics through its emphasis on personal moral standards	*HMA: moral intentionality* 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	*Syntheses* The study addresses the multifaceted ethical questions surrounding RAI and calls for a balanced approach to understanding their moral agency
Telkamp and Anderson (2022) "The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence" Journal of Practical Ethics	*Ethical AI based on diverse moral foundations* 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	*Normative ethics* This study addresses normative ethics by discussing the conflicts between consequentialism, deontology, and virtue ethics, highlighting the importance of shared moral understanding among humans	*HMA: moral understanding* 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	*Conceptualization* The study draws on the moral foundation theory of Graham et al. (2013) to illustrate why current AI frameworks do not adequately address how different organizational stakeholders make different moral judgments concerning AI

Table 8 (continued)

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
Terzis (2020) "*Onward for the freedom of others*" FAT Proceedings	*AI governance* The study critiques the current approach to AI ethics, arguing that it fails to address the fundamental issue of individual freedom. Terzis contends that ethics is an ongoing process aimed at transcending toward freedom, which cannot be captured by aggregate ethical principles or norms	*Normative ethics* The study implicitly critiques deontological and utilitarian approaches for their tendency to impose collective norms that may restrict individual moral autonomy	*AMA: explicit ethical agents* 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)	*Conceptualization* The study is criticizing that AI ethics roundtable of different stakeholders (policymakers, lobbyists, researchers, etc.) are illusionary 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 and Khomh (2023) "*The Different Faces of AI Ethics Across the World: A Principle-To-Practice Gap Analysis*" IEEE Transactions on Artificial Intelligence	*AI ethics: the principle-to-practices gap* 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 practical guidance of AI principles is required	*Normative ethics* The work implicitly engages with deontological ethics through its analysis of AI governance principles and contractarianism through its discussion of policy frameworks *Applied ethics:* The study further talks about how to apply ethics to AI strategizing, referring to AI-business alignment	*HMA: distributed agency* 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	*Syntheses* 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, fairness, security, safety, responsibility, accountability, explainability, well-being, human rights, and inclusiveness.)
Tóth et al. (2022) "*The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability*" Journal of Business Ethics	*AI robot accountability* 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	*Normative & descriptive ethics* The authors implicitly address consequentialism by emphasizing the moral intensity of decisions, considering the magnitude and distribution of consequences	*AMA: weak/strong ethical agents* 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 edict 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)	*Philosophical essay* 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

Table 8 (continued)

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
Townsend et al. (2022) "*From Pluralistic Normative Principles to Autonomous-Agent Rules*" Minds and Machines	*Institutionalized AI ethics* 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	*Applied & normative ethics* 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	*HMA: distributed agency* The authors mainly focus on distributed human agency (of developers) and oversight (of technical robustness, data governance, and so on)	*Philosophical essay* 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) "Rebuilding *'Ethics' to Govern AI: How to re-set the boundaries for the legal sector?*" ICAII Proceedings	*Limitations of AI governance* 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 interoperation 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)	*Applied ethics* The study focuses on boundaries through policy options for the legal sector, such as applying regulation	*HMA: shared responsibilities* 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	*Position paper* The study outlines legal ethics in the context of AI based on the three policy options
Vakkuri et al. (2019) *"Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study*" Computer Science	*AI ethics: ethics by design, in design, and for design* The study focuses on the implementation of AI ethics through design, proposing a 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	*Applied ethics* The study focuses on applying ethical principles in the design of AI systems	*HMA: shared responsibilities* The authors argue that there is a chain of responsibility of several stakeholders that make decisions upon the design of an AI system (internally motivated; shared responsibility; distributed agency), while to establish accountability (externally motivated), there has to be drawn connnection between all those individual's decisions and the system's actions (hybrid accountability)	*Case Study* This work is an industrial multiple case study in the healthcare sector about a baseline for ethics in AI-based software development
Vakkuri et al. (2021) "*ECCOLA—A Method for Implementing Ethically Aligned AI Systems*" The Journal of Systems & Software	*Method for ethically aligned AI* 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	*Applied ethics* 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	*HMA: shared responsibilities* Following the authors, responsibility, agency, and oversights are context parts of ECCOLA as they are based on AI ethics principles	*System design* 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

Table 8 (continued)

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
van de Poel (2020) "*Embedding Values in Artificial Intelligence (AI) Systems*" Minds and Machines	*HCAI: Embedding values into AI* 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	*Applied ethics* 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)	*HMA: shared responsibilities* The authors argue that monitoring and redesigning AI systems is important to ensure that values are kept (control agency of humans)	*Conceptualization* This work contextualizes values and what it means to embed values in AI technology
van de Voort et al. (2015) „*Refining the Ethics of Computer-made Decisions: A Classification of Moral Mediation by Ubiquitous Machines* " Ethics Information Technology	*Moral machines* 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	*Meta-ethics* This study discusses morality from a more general view	*AMA: moral responsibility* 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 argue that a moral decision requires human-like consciousness and free will, wherefore "machines do not only have morality in the sense of mediating the actions of humans but that, by making their own decisions within their relations with humans, mediate morality itself." (p. 41)	*Syntheses* This work contains a classification of ubiquitous machine-assisted moral decisions and a classification of moral decisions of machines
Vanderelst and Winfield (2018)"*The Dark Side of Ethical Robots*" AIES Proceedings	*AI design* 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)	*Normative ethics* 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)	*HMA: hybrid agency* 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 *hybrid moral agency*	*Experiment* This study builds upon an experiment with hypothetical scenarios comparing an ethical, competitive, and aggressive robot
Vanhée and Borit (2022) "*Viewpoint: Ethical By Designer—How to Grow Ethical Designers of Artificial Intelligence*" Journal of Artificial Intelligence Research	*Ethical designers of AI* 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)	*Normative ethics* 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)	*HMA: full ethical agents* 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	*Conceptualization* The study introduces a framework for designing ethical AI using with four dimensions to strengthen moral behavior of individuals: i) moral sensitivity, ii) moral judgment, iii) moral motivation, and iv) moral action/implementation

Table 8 (continued)

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
Vică et al. (2021) "*The Emperor is Naked: Moral Diplomacies and the Ethics of AI*" Információs Társadalom	*AI ethics as a buzzword* 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	*Descriptive ethics* The paper describes AI ethics as a form of diplomatic procedure, emphasizing the negotiations and compromises involved in ethical decision-making in AI *Normative ethics* The work implicitly addresses contractarianism by discussing the bureaucratization of ethics, where ethical principles are shaped and constrained by institutional procedures and agreements	*HMA: distributed agency* 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	*Philosophical essay* 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) "*Artificial Intelligence and the Ethical Use of Knowledge*" Strategica	*Aligning AI strategy with business ethics* 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	*Applied ethics* This work explores frameworks for moral decision-making in the context of aligning AI with business strategies, focusing on practical applications of ethical theory	*HMA: distributed agency* 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	*Syntheses* The paper conducts a mapping study, identifying three main approaches through which knowledge management can address ethical challenges associated with AI
Volkman & Gabriel (2023) "*AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement*" Science and Engineering Ethics	*Augment or supplant human moral reasoning* 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	*Normative ethics* The paper explicitly engages with 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	*AMA: no moral agency* The authors criticize the assumption that AI can 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	*Philosophical essay* The work proposes a modular system of multiple AI Socratic interlocutors inspired by Socratic dialogue as a means of training diverse moral traditions and pursuing moral enhancement
Walke et al. (2023) "*AI in Government: A Study on Explainability of High-Risk AI-Systems in Law Enforcement & Police Service*" WI Proceedings	*Explainable AI* 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	*Applied ethics* 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 *Normative ethics*: The paper implicitly addresses deontological ethics by emphasizing the importance of acts and laws in guiding ethical AI practices	*AMA-HMA: hybrid agency* 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	*Syntheses* This work synthesizes findings from eleven expert interviews, identifying XAI requirements for legal compliance in law enforcement and police services
Wang et al. (2020) "*Toward an Understanding of Responsible Artificial Intelligence Practices*" HICSS Proceedings	*AI ethics and corporate social responsibility* 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	*Applied ethics* The authors analyze responsible AI practices, exploring how organizations can align AI deployment with ethical principles *Normative ethics* The work implicitly references virtue ethics, linking the ethical mindset of employees to the broader ethical culture of an organization	*AMA-HMA: distributed agency* 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 extend	*Syntheses* The paper categorizes ten cases of responsible AI practices across various service industries in North America, Europe, and the UK

Table 8 (continued)

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
Weber (2020) "*Socio-Ethical Values and Legal Rules on Automated Platforms: The Quest for a Symbiotic Relationship*" Computer Law & Security Review	*Linking values to regulation for trustworthy AI systems* 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	*Applied ethics* The paper focuses on designing regulatory tools that align ethical principles with legal requirements, ensuring that AI systems are both lawful and ethically sound *Normative ethics* The paper implicitly addresses virtue ethics, referring to societal values, and cultural norms, and deontology, referring to law and regulation	*HMA: full ethical agents* The authors argue that human agency remains central in the socio-ethical dimension, emphasizing that machines cannot fully replace human moral oversight	*System design* 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) "*Ethics as Discursive Work: The Role of Ethical Framing in the Promissory Future of Data-driven Healthcare Technologies*" Science, Technology & Human Values	*Balancing regulation and evaluation within AI practices* 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)	*Applied ethics* The work emphasizes that ethics extends beyond abstract principles, addressing the practical realities of ethical decision-making in AI development	*HMA: shared responsibilities* 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	*Syntheses* This work outlines and analyzes three ethical framings and their implications for stakeholders' roles and responsibilities in AI development
Weiss et al. (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*" IFIP Proceedings	*Responsible AI design* 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)	*Applied ethics* 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	*HMA: hybrid moral agency* 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	*System design* 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) "*Ethics in the Intersection of AI and Democracy: The AIDEM Framework*" ECIS Proceedings	*AI ethics and democracy* 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	*Applied ethics* The paper applies theories of moral philosophy to analyze AI use-cases, offering a framework for resolving moral dilemmas related to AI and democracy *Meta-ethics* 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	*HMA: hybrid moral agency* 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	*Position paper* 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) "*Ethics of robotized public services: The role of robot design and its actions*" Government Information Quarterly	*Robotic ethics* 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	*Normative ethics* 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	*AMA: AI perceptions* 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	*Experiment* 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

Table 8 (continued)

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
Williams et al. (2021) "*Teacher Perspectives on How To Train Your Robot A Middle School AI and Ethics Curriculum*" AAAI Proceedings	*AI curricula* 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	*Descriptive ethics* 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	*HMA: hybrid moral agency* 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	*Case Study* 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) "*Ethics and Regulation of Artificial Intelligence*" IFIP Proceedings	*Limitations of AI governance* 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 effectively address the ethical challenges posed by AI	*Applied ethics* 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 alongside technological advancement	*HMA: shared moral responsibility* 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	*Syntheses* This work presents the current debates on regulatory systems surrounding AI based on an international conference
Wright (2020) "*AI in the Law: Towards Assessing Ethical Risks*" Big Data Proceedings	*High-level AI governance* 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	*Normative ethics* 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	*AMA: implicit/explicit ethical agents* The study draws on Winfield (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	*Syntheses* 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)

Table 8 (continued)

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
Wu et al. (2020) "*Ethical Principles and Governance Technology Development of AI in China*" Engineering	*HCAI for sustainable development of AI* 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	*Applied ethics* The study describes the ethical principles guiding AI governance in China, focusing on the country's efforts to balance technological innovation with ethical responsibility	*HMA: shared moral responsibility* 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	*Syntheses* This work describes the efforts of AI governance in China and the preliminary outcomes
Yu et al. (2018) "*Building Ethics into Artificial Intelligence*" JCAI Proceedings	*Ethical decision-making for AI governance* 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	*Normative ethics* 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	*HMA: Limited human agency:* 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	*Syntheses* This work summarizes psychological, social, and legal discussions on AI governance and proposes a taxonomy of AI governance techniques
Zebrowski (2021)"*Artificial Instinct: Lem's Robots as a Model Case for AI*" Pro-Fil	*More nuanced models for AI ethics* 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	*Meta-ethics* The study explores the meta-ethical implications of AI, combining metaphysical and ethical inquiries to challenge the anthropocentric biases in current AI ethics frameworks	*HMA vs. AMA:* 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," which simulates basic human instincts rather than higher cognitive functions	*Philosophical essay* 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 and Zhang (2023) *"Ethics and Governance of Trustworthy Medical Artificial Intelligence*" BMC Medical Informatics and Decision Making	*AI governance in healthcare* 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	*Normative ethics* 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	*HMA: full ethical agents* 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	*Syntheses* 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)

Table 8 (continued)

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
Zhang et al. (2021) "*Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers*" Journal of Artificial Intelligence Research	*Global AI governance* This paper builds on the definition of AI governance as the global norms and policies that ensure the beneficial design and use of AI. The study highlights the need for a coordinated international effort to develop governance frameworks that can effectively manage the ethical challenges posed by AI	*Descriptive ethics* The work provides a descriptive analysis of the ethical challenges in AI governance based on insights from machine learning researchers. It emphasizes the importance of understanding the ethical concerns of those who are directly involved in AI development	*HMA: hybrid moral agency* The study underscores the expectations of AI researchers for executives and policymakers to develop trustworthy AI systems and establish appropriate governance frameworks. It highlights the shared responsibility for ensuring that AI systems are ethically designed and implemented	*Syntheses* This work compares the consensus and disagreements of AI researchers regarding the ethics and governance of AI

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

Appendix 6. 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., 2021; Minkkinen & 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, 2015a, 2015b, 2015c; 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, 2015a, 2015b, 2015c; 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; Koebis 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, 2022; 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, 2020; Vică 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

Table 9 (continued)

Category	Meta Information	Statistics	Articles
Publication Type	Journal	N = (92/151) 60,93%	Allen et al., 2000; Arnold & Scheutz, 2018; Ashrafian, 2015a, 2015b, 2015c, 2015b, 2015c, 2015a, 2015b, 2015c, 2015b, 2015c; 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; Koebis 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

Table 9 (continued)

Category	Meta Information	Statistics	Articles
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; Kornilaev, 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, 2015a, 2015b, 2015c, 2015b, 2015c, 2015a, 2015b, 2015c, 2015b, 2015c; 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 & Çürüklü, 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; Koebis et al., 2021; Kučera, 2019; Lara, 2021; Li, 2021; Mabaso, 2021; Mäntymäki et al., 2023; Mayer et al., 2021; 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; Minkkinen & 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., 2021; 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

Table 9 (continued)

Category	Meta Information	Statistics	Articles
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., 2021; 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, 2015a, 2015b, 2015c, 2015b, 2015c, 2015a, 2015b, 2015c, 2015b, 2015c; 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; Koebis 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; Saariluoma & 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; Vică 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

Table 9 (continued)

Category	Meta Information	Statistics	Articles
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; Koebis 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, 2015a, 2015b, 2015c; 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., 2021; Townsend et al., 2022
	No specific context	N = (98/151) 64,9%	Arnold & Scheutz, 2018; Ashrafian, 2015a, 2015b, 2015c; Auld et al., 2022; Badea & Artus, 2022; Bankins & 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; Ghatar 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; Kornilaev, 2021; Kowch, 2019; Li, 2021; Lim & Kwon, 2021; Lloyd, 1985; Mabaso, 2021; Manna & Nath, 2021; Martin, 2019; Martinho et al., 2021; Millar, 2016; Minkkinen & 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

Table 9 (continued)

Category	Meta Information	Statistics	Articles
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, 2015a, 2015b, 2015c; Bankins & 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; Ghatar 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; Minkkinen & 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, 2015a, 2015b, 2015c; Bryson, 2018; Constantinescu et al., 2021; Cook, 2023; Erman & Furendal, 2022; Etzioni & Etzioni, 2017; Figueras et al., 2022; Ha, 2022; Hauer, 2020; Kornilaev, 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., 2021; 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 7. 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 *(deductive)*	• 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 categorial 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 *(deductive)*	• 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, ones 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 *(deductive)*	• 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 *(deductive)*	• 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 *(deductive)*	• 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 & Nagel, 1974), Normative Democratic Theory mainly based on Rawl (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 *(inductive)*	• 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 *(inductive)*	• 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 *(inductive)*	• 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 seamful design must be accompanied by stringent accountability measures." (Susser, 2019, p. 407)

Table 10 (continued)

Category	Explanation	Codes	Coding Rules	Example
Descriptive ethics (Ethical theory) Context	Is about empirically describing people's moral and ethical believes and behaviors	Trolley problem *(inductive)*	• 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 *(inductive)*	• 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 Allen 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 *(inductive)*	• 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. 15,678)
		Business ethics *(deductive)*	• 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 oversights, 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 intension 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., 2020, 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)

Table 10 (continued)

Category	Explanation	Codes	Coding Rules	Example
Presence of ethical theory	Identifies whether the ethical theory is explicitly mentioned or implicitly applied in the research	Explicit *(deductive)*	• 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 *(deductive)*	• 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) *(deductive)*	• 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) *(deductive)*	• 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 *(inductive)*	• 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 *(inductive)*	• 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 patiency, 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 patiency 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, and thus being able to take responsible for those actions	Explicit ethical agents *(deductive)*	• 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 ethical agents. Average adult human beings are an instance of this type of agent." (Boyles, 2018, p. 195)
		Implicit ethical agents *(deductive)*	• 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 *(deductive)*	• 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 *(inductive)*	• 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 *(inductive)*	• 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)

Table 10 (continued)

Category	Explanation	Codes	Coding Rules	Example
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 interaction takes place	Regulating human-AI interaction *(inductive)*	• 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 Emoters in the data that the models were trained on…" (Ong, 2021, p. 5)
		Designing human-AI interaction *(inductive)*	• 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 *(inductive)*	• 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 *(inductive)*	• 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. 15,678)
		Theorizing/Researching upon human-AI interaction *(inductive)*	• 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)

Table 10 (continued)

Category	Explanation	Codes	Coding Rules	Example
Research Method, Design, Approach, or Philosophy	Describes the underlying methodological approaches	Syntheses (MacInnis, 2011) *(deductive)*	• Research method focusing on the identification of relevant material and the aggregation and classification/clustering of the material • Summarizes explicit and tangible material (facts, principles, law, etc.) • Related terms: summary, typology, mapping studies	e.g., Tasioulas (2018), van der Voort et al. (2015)
		System Design (Siau & Rossi, 2011) *(deductive)*	• 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 & Scheutz (2018)
		Conceptualization (MacInnis, 2011) *(deductive)*	• 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., Vanhée and Borit (2022), Shneiderman (2020)
		Philosophical Essay (Creswell & Creswell, 2014, pp. 5–11) *(deductive)*	• Philosophical worldviews • Summarizes implicit and intangible thoughts, opinions, and beliefs • Related terms: academic essay	e.g., Susser (2019), Swanepoel (2021)
		Position Paper (Creswell & Creswell, 2014, pp. 5–11) *(deductive)*	• 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 & Creswell, 2014, p. 13) *(deductive)*	• 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 & Creswell, 2014, pp. 13, 170–174) *(deductive)*	• 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 & Creswell, 2014, pp. 155–173) *(deductive)*	• Based on an online questionnaire with measures and scales • Statistical or descriptive analysis	e.g., Khan et al. (2022)
		Quantitative Text Analysis (Creswell & Creswell, 2014, pp. 46–47) *(deductive)*	• Analyzing textual content from various sources (e.g., social media, documents) • Statistical or descriptive analysis	e.g., Neubert & Montanez (2020)
		Qualitative Content Analysis (Creswell & Creswell, 2014, pp. 186, 197–201) *(deductive)*	• 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 and Zamani (2022)
		Interviews (Creswell & Creswell, 2014, p. 190) *(deductive)*	• 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 and Pan (2020), Stahl et al. (2022)

Appendix 8. 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

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 9. Author Citation Network

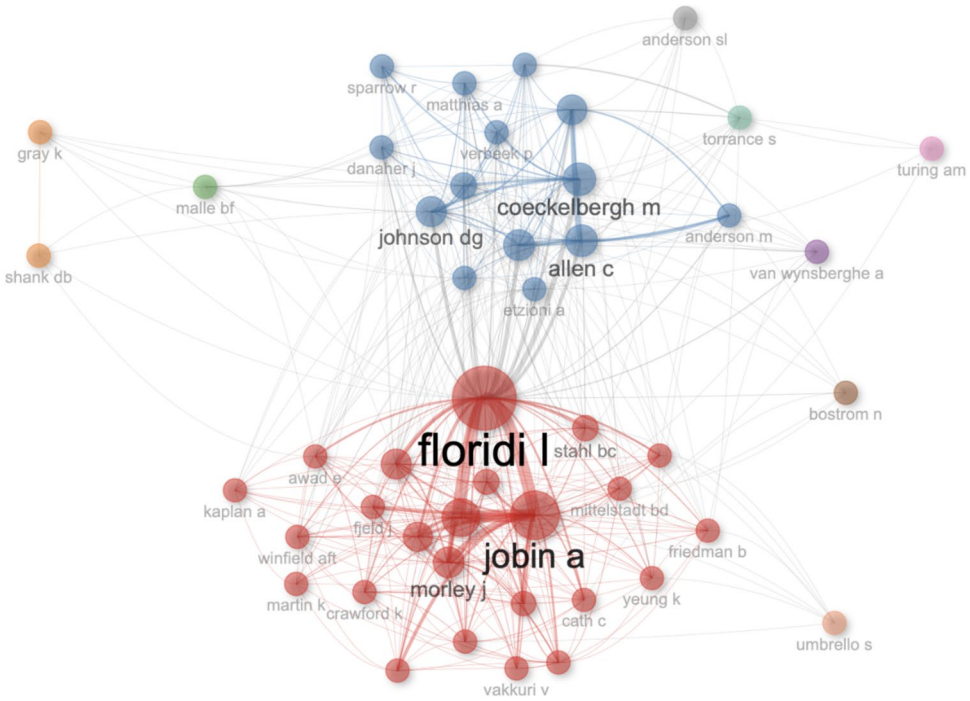


Fig. 5 Zoom-in in the author citation network (n = 50 top cited authors)

Appendix 10. Distribution of Articles over Years

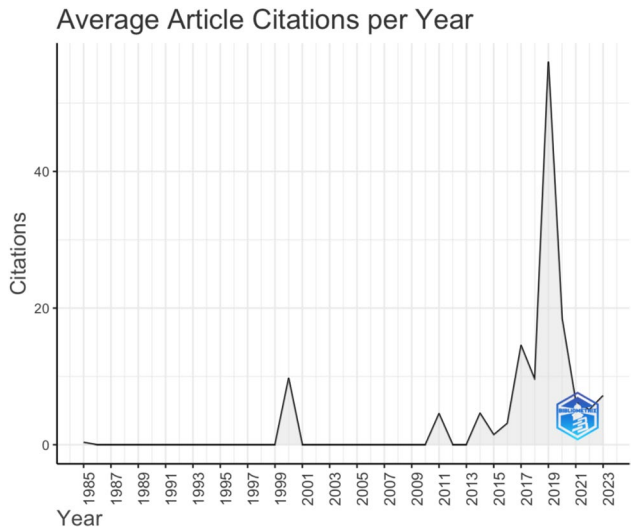


Fig. 6 Average citations of the citation network per year (n = 137)

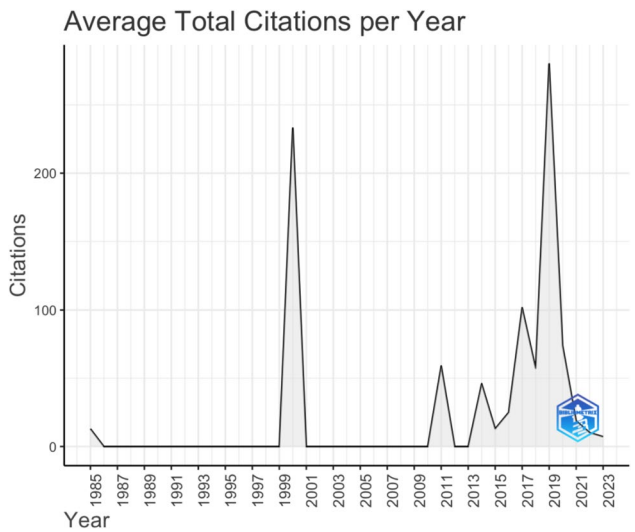


Fig. 7 Total citations of the citation network per year (n = 137)

Appendix 11. Most Cited Publication Outlets

We further employed bibliographic coupling analysis (node size = 20). We analyzed the level of articles as seen in Fig. 8 using the Walktrap clustering algorithm.



Fig. 8 Most cited publication outlets (n = 20)

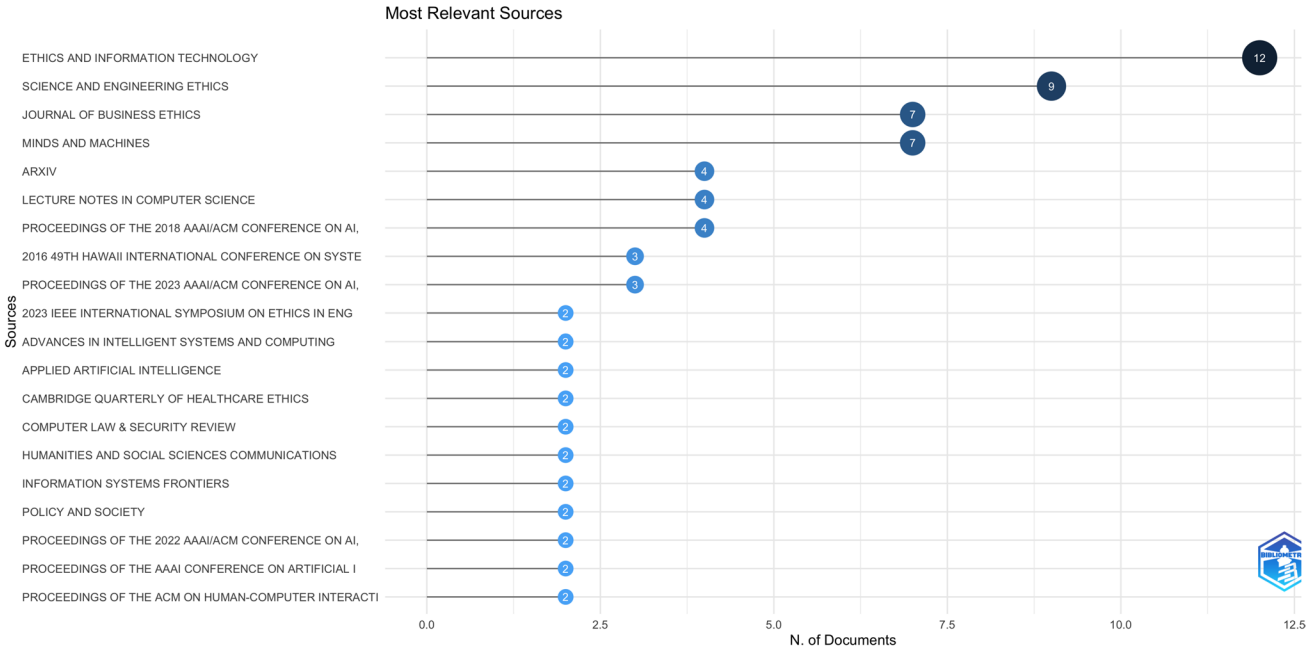


Fig. 9 Most relevant sources (n = 137)

Appendix 12. Author Citations based on Authors Countries

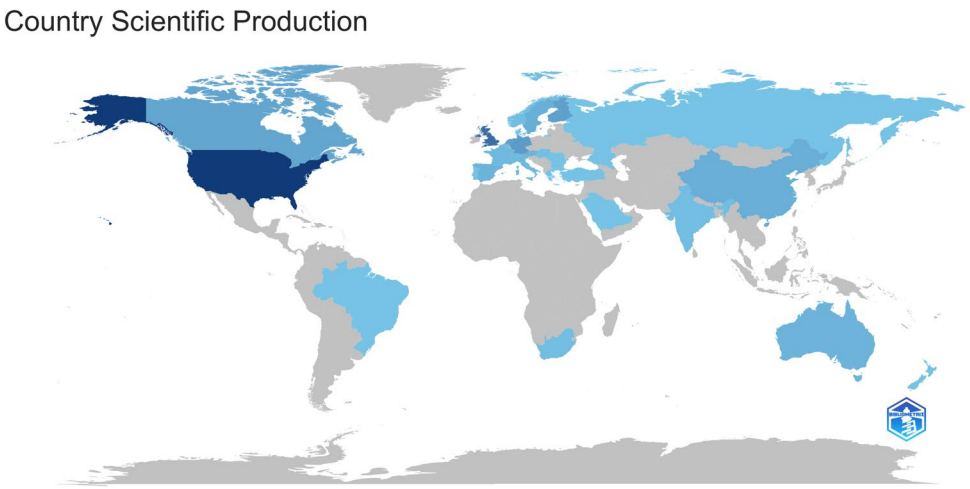


Fig. 10 Country distribution of authors (n = 6,523)

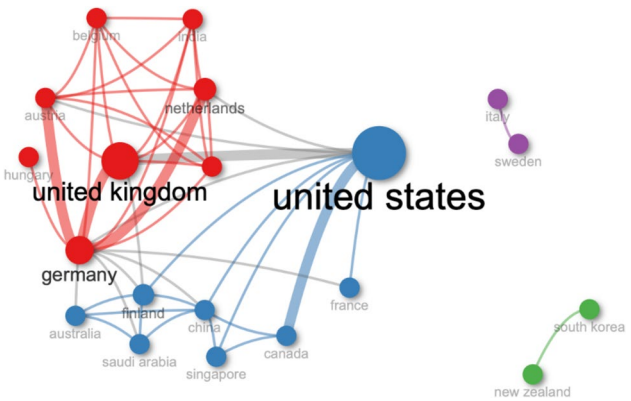


Fig. 11 Collaboration network of the final sample's first authors (n = 137)

Appendix 13. 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 14. 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 (*the AI is directly influenced by the developer's environment*). This can further lead to an indirect influence of the human user, when the AI technology is designed to actively avoid misusage *(the human user is indirectly influenced by the AI's design)*	Deontology	e.g., Boyles (2018), Génova et al. (2023), Hauer (2020), Manna and 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 (*the human user is influenced by the organizational environment*). There can be carrot and stick mechanism that enforce norm-compliance	Consequentialism	e.g., Bankins and Formosa (2023), Millar (2016), Pruski (2023), Tóth 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 (*the AI is influenced by the regulatory environment).* As the AI technology itself cannot be judged prosecution or carrot and stick mechanisms are directed at AI providers	Contractarianism	e.g., Gómez de Ágreda (2020), Nieder-man and 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

Table 13 (continued)

Mechanism	Driving entities	Ethical Theory	Authors	Research Gaps in the Context of Descriptive ethics
Teaching ethical norms combined with AI knowledge within AI education enhances human users' competencies, skills and knowledge (AI 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	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 (*the human user is influenced by the societal environment*)	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
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 (*human user's feedback influences the AI's underlying processing models and outputs*)		e.g., Etzioni and Etzioni (2017), Greene et al. (2019), Roberts et al. (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 (*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*)	Applied ethics	e.g. Etzioni and Etzioni (2017), Dodig Crnkovic and 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 (*human perceptions influence the AI*)	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

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

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.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received.

Data Availability Not applicable.

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.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

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 (URI: https://hdl.handle.net/10125/64389 978-0-9981331-3-3).

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine, 28*(4). https://doi.org/10.1609/aimag.v28i4.2065

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. (2015). AIonAI: A Humanitarian Law of Artificial Intelligence and Robotics. *Science and Engineering Ethics, 21*(1), 1. https://doi.org/10.1007/s11948-013-9513-9

Ashrafian, H. (2015). 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. (2015). Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights. *Science and Engineering Ethics, 21*(2), 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., spsampsps Artus, G. (2022). Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents. In M. Bramer spsampsps 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)*. https://aisel.aisnet.org/acis2021/52

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). 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

Burema, D., Debowski-Weimann, N., von Janowski, A., Grabowski, J., Maftei, 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 Chemsitry. *Scientometrics, 22*(1), 155–205. https://doi.org/10.1007/BF02019280

Carneiro, D., spsampsps Veloso, P. (2022). Ethics, transparency, fairness and the responsibility of artificial intelligence. *New Trends in Disruptive Technologies.* https://doi.org/10.1007/978-3-030-87687-6_12

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., spsampsps 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

Creswell, J. W., & Creswell, J. D. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

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). Available at: https://aisel.aisnet.org/sjis/vol34/iss2/6

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

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

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*, 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. https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/4

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

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55–130.

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*. https://hdl.handle.net/10125/59651

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, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions? *Society, 57*(4), 425–433.

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 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

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*. IRIS Association. IRIS. https://aisel.aisnet.org/iris2021/5

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*. 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*. http://arxiv.org/abs/2207.01493. Accessed 14 Mar 2024.

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), 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

Kornilaev, 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

Krügel, 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. https://doi.org/10.18267/pr.2019.dvo.2316.0

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, spsampsps 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. https://doi.org/10.1145/3463677.3463733

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

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 Books.

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*. https://aisel.aisnet.org/ecis2021_rp/27

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

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%3c433::AID-ASI11%3e3.0.CO;2-Q

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). *A proposal for the Dartmouth summer research project on artificial intelligence*, August 31, 1955. AI Mag. 27, 12.

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. *Candinavian 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, 21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80

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

Nozick, R., & Nagel, T. (1974). *Anarchy, State, and Utopia* (Vol. 5038). Basic Books.

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.

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. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/15

Polyviou, A., & Zamani, E. D. (2022). Towards Europe's AI Strategy: Desires & Realities. *Thirtieth European Conference on Information.* Timișoara, Romania Systems. https://aisel.aisnet.org/ecis2022_rp/48

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., Cowls, 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. Accessed 11 Nov 2023.

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)*. Printed in the United States of America by Vail-Ballou Press, Binghamton, New York. Library of Congress Cataloging-in-Publication Data.

Saariluoma, P., spsampsps Leikas, J. (2020). Designing Ethical AI in the Shadow of Hume's Guillotine. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, spsampsps 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*. 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) and the 6th Middle East & North Africa Conference on digital Information Systems (MENACIS).*

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.* Kristiansand, Norway. Systems. https://aisel.aisnet.org/ecis2023_rp/368

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., & Rossi, M. (2011). Evaluation techniques for systems analysis and design modelling methods–a review and comparative analysis. *Information Systems Journal, 21*(3), 249–268.

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 Metaphyscis, 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., & 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

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

Strawson, P. F. (2008). Freedom and resentment. *Freedom and Resentment and Other Essays*.

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

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., 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

Vakkuri, V., Kemell, K.-K., spsampsps Abrahamsson, P. (2019). Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study. In X. Franch, T. Männistö, spsampsps 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

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). https://doi.org/10.1007/s10796-020-10096-3

Vică, C., Voinea, C., & Uszkai, R. (2021). The Emperor is Naked: Moral Diplomacies and the Ethics of AI. *Információs Társadalom, XX, I*(2), 83–96.

Vidu, C., Zbuchea, A., Mocanu, R., & Pinzaru, F. (2020). Artificial Intelligence and the Ethical Use if 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)*. https://aisel.aisnet.org/wi2023/76

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., spsampsps 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, spsampsps 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*. https://aisel.aisnet.org/ecis2023_rp/321

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., Schmidthuber, 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

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics, 2*(1), 46–48. https://doi.org/10.1038/s41928-019-0213-6

Wong, A. (2021). Ethics and Regulation of Artificial Intelligence. In E. Mercier-Laurent, M. Ö. Kayalica, spsampsps 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. *IEEE International Conference on Big Data (Big Data), 2020*, 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, 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

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

Zupic, I., & Čater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods, 18*(3), 429–472. https://doi.org/10.1177/1094428114562629

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Teresa Hammerschmidt (neè Heyder) is an affiliate member of the Department of Information Systems and Social Networks at the University of Bamberg. She completed her Ph.D. in Information Systems from the University of Bamberg in 2025. Her research primarily investigates the ethical management of human-AI interaction within organizations, with a focus on sociotechnical system theory.

Alina Hafner is a Ph.D. candidate at the Chair of Information Systems and Business Process Management at the Technical University of Munich. She received her MS in Information Systems from the University of Bamberg (2023). Her research focuses on (Generative) AI, Business Process Management as well as AI ethics.

Katharina Stolz completed her doctoral studies at the University of Stuttgart in 2024 at the Chair of Innovation and Service Management. Her research focusses on ethical management of human-AI interactions and innovation structures, with a particular focus on the luxury industry.

Nina Passlack is a doctoral candidate at the Faculty of Information Systems and Applied Computer Sciences at the University of Bamberg, Germany. She holds a Master of Science in International Information Systems Management from the University of Bamberg (2021). Her research primarily explores socio-technical issues, with a particular emphasis on responsibility attribution in human-AI interactions and the competencies necessary for the responsible use of AI.

Karl-Heinz Gerholz is a Professor for Business and Human Resource Education at the University of Bamberg. His research fields are digital learning and AI in Teacher Education and Vocational Education, Educational Management, Leadership Development and Vocational and Education Training on a policy level.

Oliver Posegga is a professor at the Department of Information Systems and Social Networks at the University of Bamberg, Germany. His research focuses on social networks, digital collaboration, and the sociotechnical dynamics of hybrid media systems as well as the ethical management of AI, AI literacy, and digital responsibility.

Paper III

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 (BIT)*, 1–30.

https://doi.org/10.1080/0144929X.2024.2431050

Paper IV

Passlack, N., Hammerschmidt, T., & Posegga, O. (2025). How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice. *American Conference on Information Systems 2025 Proceedings (AMCIS)*. 4

https://aisel.aisnet.org/amcis2025/is_leader/is_leader/4/

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2025 Proceedings

Americas Conference on Information Systems (AMCIS)

August 2025

How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice

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

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

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

Follow this and additional works at: https://aisel.aisnet.org/amcis2025

Recommended Citation

Passlack, Nina Katharina; Hammerschmidt, Teresa; and Posegga, Oliver, "How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice" (2025). *AMCIS 2025 Proceedings*. 4.
https://aisel.aisnet.org/amcis2025/is_leader/is_leader/4

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2025 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice

Completed Research Full Paper

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

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

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

Abstract

As algorithmic advice becomes increasingly integral to leaders' decision-making, concerns about moral responsibility grow more pressing. While algorithms are often praised for their objectivity and efficiency, leaders remain ultimately accountable for their decisions, even when relying on them. This study builds on prior research, which reveals a tendency for individuals to claim greater responsibility for successful outcomes while shifting blame for failures when following algorithmic advice, by examining the role of moral identity in shaping this dynamic. Through vignette experiments, the research contributes to Information Systems (IS) research by providing empirical evidence on how leaders' moral identity influences their responsibility attribution in algorithmic advised decision-making. By doing so, the study also offers practical implications for organizations seeking to foster ethical leadership.

Keywords

Morality, algorithmic advice, moral agency, responsibility attribution, responsible human-AI interaction.

Introduction

As algorithms can efficiently process vast quantities of data to predict outcomes, leaders increasingly rely on algorithmic advice when making organizational decisions (Gunaratne et al., 2018; van der Vlist, 2016). Algorithmic advice offers certain advantages for leaders' decisions, particularly since algorithms are considered to have no emotionality (Giroux et al., 2022) and are more objective due to their norm-compliance and by being able to take more data into account, as humans can (Sundar & Kim, 2019). However, leaders are also aware of the fact that in specific types of decisions, for instance, those that affect employees' personal lives, their human empathy is required (Passlack et al., 2024). Thus, according to prior research, it is important that the final decision remains with humans (Gerdes, 2022; Nevanperä et al., 2021). This raises questions on how algorithmic advice influences leaders' responsibility perceptions, particularly when outcomes are not as intended.

Although leaders are remaining accountable for their decisions in practice (Horneber & Laumer, 2023), regardless of whether they received algorithmic advice, there is an increased concern regarding the extent to which individuals perceive themselves as responsible for outcomes when relying on algorithmic advice (e.g., Bleher & Braun, 2022; Strich et al., 2021). Research has highlighted that leaders tend to attribute greater responsibility to themselves for successful outcomes, while attributing less responsibility for

unsuccessful ones (e.g., Mayer et al., 2020; Passlack et al., 2024). This discrepancy creates responsibility gaps, underscoring the need for a deeper examination of the factors that shape leaders' responsibility attribution when receiving algorithmic advice in their daily decision-making (McGuire & De Cremer, 2023). As the work of Passlack et al. (2023) outlined, leaders' morality may play a crucial role in how they perceive and navigate these responsibilities, as moral considerations are integral to decision-making processes in the context of algorithmic reliance.

Leaders' morality can be seen as inherent moral principles and an intrinsic moral motivation, which refers to the extent to which a person considers moral character to be a core part of their self-identity (Hardy & Carlo, 2011). This moral character guides leaders in their decision-making, influencing their perceptions of responsibility (Heyder et al., 2023). While there is no universal agreement on the precise definition of one's morality, some perspectives emphasize the importance of considering moral agency when attributing responsibility (Constantinescu et al., 2021; Hammerschmidt et al., 2025). Moral agency refers to the capacity of individuals to pursue their actions through their intentional reflection on what matters morally (Barnes, 1999). Thus, in this paper, we aim to examine how leaders' moral identity – guiding their moral agency – influences the extent to which they assume responsibility for positive and negative outcomes of their decisions, which they made with the help of algorithmic advice. The research question (RQ) guiding this investigation is:

How does leaders' morality impact their self-attribution of responsibility when receiving algorithmic advice in case of failure and success?

Upon conducting vignette experiments, we investigated how leaders' morality influences their responsibility attribution toward themselves and the algorithmic advice when making decisions in an organizational context by relying on algorithmic advice. We conducted vignette experiments with four different decision-making tasks that were brought forward by Passlack et al. (2024). While prior research exists that investigated how self-attribution of responsibility differs in cases of failure and success, we are interested in the impact of leaders' morality on their self-attribution of responsibility behavior.

This study aims to contribute to current Information Systems (IS) research by providing empirical evidence on the effects of a leader's morality on self-attribution of responsibility for decision-making outcomes when receiving algorithmic advice. It is unique in its sample, focusing on leaders who hold significant responsibilities within organizations, including responsibilities for people. Decisions in organizational contexts are highly ethical in nature and, therefore, require a strong sense of morality to guarantee that leaders also assume responsibility for misguided decisions. Second, to the best of the authors' knowledge, there is currently little empirical research on moral identity and its impact on perceived responsibility in the context of human-AI interaction. The increasing autonomy of artificial intelligence (AI) raises growing questions about responsibility, and this study is particularly interested in how individuals' morality influences these perceptions. While prior research suggests that people tend to take on less responsibility for decisions with negative outcomes (e.g., Passlack et al., 2024; Strich et al., 2021), this paper provides evidence on how this can be affected by leaders' morality. In doing so, the findings are expected to contribute not only to IS research but also to inform practitioners about how professionals' morality might influence responsible decision-making in organizations.

Theoretical Background & Hypotheses Development

The following sections provide theoretical foundations on the psychological factors behind responsibility attribution and its connections to morality, in particular, moral identity. It then follows the derivation of the hypotheses resulting in our research model in Figure 1.

Responsibility Attribution and Morality

Human nature compels us to seek explanations for behaviors and decisions, both our own and those of others, shaping our emotional responses and subsequent actions (Weiner, 1985b). For instance, a leader who attributes a team's success to their own strategic acumen and effective guidance is likely to experience a profound sense of pride and accomplishment. Conversely, if the leader perceives the success as predominantly driven by external, uncontrollable factors, their sense of satisfaction may be reduced. Thus, this attributional inclination plays a crucial role in shaping leadership perceptions on responsibility, but

also leaders' psychological states. Responsibility fundamentally concerns the "what"— that is, the specific duties and tasks assigned to individuals, requiring them to act and make decisions in a responsible manner (Horneber & Laumer, 2023). When discussing moral responsibility, the focus is on whether leaders have the ability and willingness to acknowledge, act upon, and uphold responsibility (Hammerschmidt et al., 2025). This relates to their agency and their underlying sense of what is right or wrong (Novelli et al., 2024).

Weiner's (1985a) attribution theory consolidates earlier research, including Heider's (1958) work on attribution theory and Rotter's (1966) discussion on locus of control, to offer a structured framework for understanding how individuals perceive causes and, thus, attribute responsibilities. The framework is built on three key dimensions: *causality*, *controllability*, and *intentionality*. The causality dimension distinguishes between internal factors, such as an individual's effort and personality traits, and external factors, such as external assistance or unforeseen circumstances. Controllability pertains to the degree to which an individual is perceived to exert control over their actions and subsequent outcomes. Finally, intentionality considers whether a given behavior is perceived as deliberate or incidental. For instance, when a leader's decision is influenced by algorithmic advice, perceptions of controllability may decline, particularly if the algorithm is perceived as opaque or unpredictable (Passlack et al., 2024). Similarly, if a leader's decision leads to unintended negative consequences, their attribution of responsibility may be mitigated due to a lack of malicious intent (Kegel & Ghanem, 2024).

While all three dimensions are expected to influence individuals' causal attributions, in this study, we focus on the *causality* dimension in that we examine how leaders perceive causes for outcomes resulting from their decision-making when they receive algorithmic advice. This allows us to investigate leaders' self-attribution of responsibility. More precisely, when leaders perceive the causes for the performance to lie in themselves, meaning the performance was due to their internal factors, we refer to this as *self-attribution of responsibility*. Conversely, if leaders perceive the algorithmic advice as the primary determinant of the decision outcome, this reflects responsibility attribution toward external factors, thus reducing self-attribution of responsibility.

Building on Weiner's (1985a) framework, individuals' causal ascriptions hinge on their prevailing moral perceptions (Passlack et al., 2024). Moral perceptions are expected to guide leaders' decision-making in that it is linked to their moral agency. Moral agency denotes an individual's capacity to make moral judgments and act upon them based on free will and reflection (Barnes, 1999), which shapes how individuals perceive and execute decisions (Black & Reynolds, 2016). Thus, the self-attribution of responsibility necessitates considerations of moral agency (Constantinescu et al., 2021). This is because morality is expected to shape both emotions and reasoning, which in turn influence an individual's judgment and, consequently, their behavior (Schwartz, 2016). For leaders, having a strong moral agency means they are intrinsically motivated to make decisions that align with ethical principles. As a result, this strong moral agency is associated with a greater willingness to take responsibility for the consequences of their actions (Heyder et al., 2023; Reynolds & Ceranic, 2007).

Leaders' morality, in this work, is operationalized through the concept of moral identity – the degree to which morality is central to an individual's self-concept (Hardy & Carlo, 2011). Moral identity concerns moral perceptions of individuals based on their inherent ethical understanding of their beliefs to be right or wrong (Reynolds & Ceranic, 2007; Tóth et al., 2022). Hence, moral identity drives how humans make decisions and judgments based on their understanding of responsibility (Black & Reynolds, 2016). Having a high moral identity implies that people reflect on their mistakes, act honestly, treat others in a friendly way, and do not cheat (Black & Reynolds, 2016). Thus, it is expected to be central to a leader's responsibility attribution for failures in algorithmic advised decision-making.

Given the increasing integration of algorithms in decision-making processes in practice, leaders' morality becomes a crucial factor in mitigating responsibility gaps and unethical decisions that can arise when leaders feel less responsible for their decision-making outcomes (Heyder et al., 2023). While some scholars conceptualize moral identity as a trait-like attribute in which morality is central to an individual's self-concept (e.g., Frimer & Walker, 2009), thus being a relatively stable trait, other perspectives frame it as a more dynamic and context-dependent construct, shaped by deliberative and situational processes (e.g., Lapsley & Narvaez, 2004). Thus, the practical relevance of moral identity lies in the ability of organizations to create structures and environments that support and activate moral behavior. This ensures that leaders – who are formally entrusted with responsibility – also assume accountability for outcomes (Passlack et al.,

2024). The following section explores how leaders' morality, operationalized through their moral identity, is more inclined to uphold responsibility, even when decision-making is advised by algorithms.

Hypothesis Development

For testing to what extent leaders' moral identity influences their self-attribution of responsibility in case of failure and success in algorithmic advised decision-making, we derive hypotheses to test the following research model, presented in Figure 1.

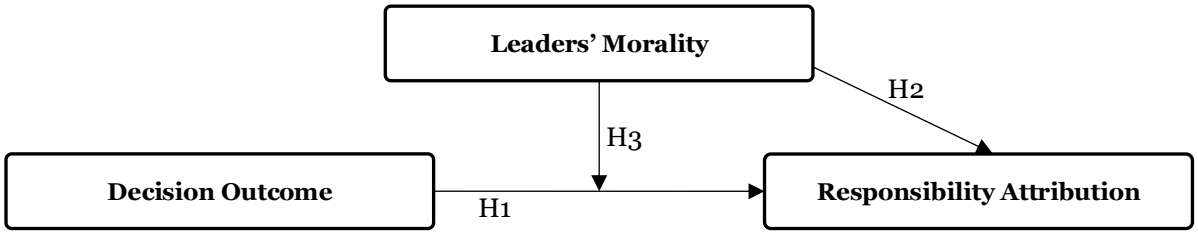


Figure 1. Research Model

Prior research shows that individuals are more likely to attribute responsibility to themselves when outcomes are successful compared to situations where they experience failure. This pattern has been observed in both human-human interactions (e.g., Greene et al., 2022) and human interactions with AI (e.g., Passlack et al., 2024). The tendency to claim responsibility for positive results while distancing oneself from negative outcomes is driven by fundamental psychological and social motives. One key explanation lies in the desire to maintain a favorable self-image and avoid the discomfort associated with failure (Nicholls, 1975). Additionally, responsibility attributions play a crucial role in social reputation management as individuals are motivated to present themselves positively to others, as social standing and perceived competence influence future opportunities and relationships (e.g., Rowe et al., 2009). This pattern aligns with Weiner's (1985a), which states that the attribution of causes highly affects individuals' psychological state, thus their feelings. By attributing success to internal factors, individuals experience pride and heightened self-worth, whereas attributing failure to external factors reduces feelings of guilt, shame, and incompetence (Harvey et al., 1981). The interplay between self-protection and social image preservation explains why responsibility attribution is asymmetrical across positive and negative outcomes. Given these tendencies, we hypothesize that:

H1. Self-attribution of responsibility is higher among leaders facing success than those facing failure.

The impact of morality on human behavior has been explored, drawing on foundational ethical theories that highlight the significance of moral agency in shaping individual actions (Hammerschmidt et al., 2025; Novelli et al., 2024). As algorithms are increasingly built upon AI models that operate (semi-)autonomously and exhibit behaviors that resemble human cognition (Berente et al., 2021), individuals may attribute a form of agency to AI (Müller, 2021). This phenomenon raises critical concerns about responsibility diffusion, where decision-makers may shift accountability onto AI, perceiving it as an independent actor rather than a tool (O'Neill et al., 2020). For instance, research shows that when individuals work alongside AI, they tend to be held less accountable for the same actions compared to those acting alone (Shank & DeSanti, 2018). This suggests that the moral evaluation of advanced technologies often focuses more on reducing human blaming rather than assigning blame to the machines themselves (Shank & DeSanti, 2018). Morality might emerge as a key determinant of the responsibility attribution process, as it encapsulates individuals' deeply ingrained ethical principles and their commitment to upholding moral responsibility in decision-making (Reynolds & Ceranic, 2007; Tóth et al., 2022). Thus, we expect that individuals with a strong moral identity are more inclined to acknowledge their role in decision-making, even when algorithms play a role in shaping outcomes. This internalized moral identity might reduce the tendency to deflect blame onto external entities, reinforcing self-attribution in cases of both success and failure. Thus, we assume that leaders with higher moral identity are likely to resist the diffusion of responsibility, maintaining ownership over their decisions even when receiving algorithmic advice. Thus, we hypothesize that:

H2. Leaders' moral identity is positively associated with their self-attribution of responsibility.

While moral identity directly influences leaders' tendency to attribute responsibility to themselves (H2), it is also expected to moderate the relationship between decision-making outcomes and self-attribution of responsibility (H3). Prior research indicates that individuals generally attribute successes to internal factors (e.g., their own competence and decision-making abilities) while externalizing responsibility for failures to protect their self-image and reputation (e.g., Passlack et al., 2024). However, leaders with a strong moral identity are guided by ethical self-concept and integrity, which may reduce the tendency to deflect responsibility when outcomes are unfavorable (Hardy & Carlo, 2011; Reynolds & Ceranic, 2007). Thus, moral identity might foster a sense of accountability that overrides self-serving attribution biases, making leaders more consistent in taking responsibility regardless of whether the outcome can be claimed as success or failure. Thus, moral identity is expected to be a moderating factor that shapes how decision outcomes influence self-attribution of responsibility. Specifically, leaders with lower moral identity are expected to be more likely to exhibit attribution asymmetry – claiming responsibility for success but shifting blame in failure scenarios – whereas leaders with high moral identity are expected to remain more stable in their self-attribution of responsibility, even when the outcome is unfavorable. Therefore, we hypothesize:

H3. The relationship between responsibility attribution and decision outcome is moderated by leaders' moral identity in that moral identity is positively associated with self-attribution of responsibility.

Method

This work builds upon the vignette experiment published in the work of Passlack et al. (2024). The experiment consists of four decision-making tasks that are advised by an algorithm, while managers are told to receive advice from a human consultant serving as a control group. For this work, we only focus on the group that received algorithmic advice. While prior research focused on the differences in self-attribution of responsibility when receiving advice from humans or algorithms (Passlack et al., 2024), this paper extends those findings by investigating to what extent the leaders' moral identity moderates this effect. The experiment was realized via an online survey using SoSci Survey (https://www.soscisurvey.de)

Vignette experiments are perceived as a suitable method as they allow us to investigate scenarios close to real-world ones, but ensure that all participants face the scenarios under the same conditions (Atzmüller & Steiner, 2010). This is why vignette studies have strong internal validity, meaning they effectively ensure that participants' responses are genuinely influenced by the presented scenario rather than external factors (Steiner et al., 2016) and allow for making statistical inferences in contrast to qualitative interviews. Unlike traditional surveys, they actively engage participants in meaningful actions, providing more authentic and reliable insights (Atzmüller & Steiner, 2010). The following sections will outline the procedure of the vignette experiment, provide details on the participants, outline the measures used to test the proposed variables, and finally, detail the data analysis.

Procedure

The procedure of the experiment was as follows. At first, the leaders were asked to provide information on their demographic data and their positions to double-check whether their position in the organization was linked to leadership. If people were said to be employees, for instance, they were excluded from the final sample. After a brief introduction informing them that they were managers of a company facing different decision-making scenarios, the participants were presented with their first decision-making task (see Figure 2 for an example). They received algorithmic advice, which included information about the algorithm's database and the choice the algorithm would recommend. Participants were free to either follow the advice or make their own decision. After making their decision, participants received feedback on their decision that was randomly framed as either a success or a failure. Next, participants completed the responsibility attribution scale (see the "Measures" section). This process was repeated four times, ensuring that each participant encountered all the decision-making tasks. Finally, participants were asked to complete the moral identity scale (see the "Measures" section). Figure 2 illustrates a task as an example.

Your marketing agency is seeking a new marketing manager. Below is the job description outlining the necessary skills. Additionally, you've received three applications; information on their CVs is listed below. Now, you must select the candidate you want to hire.



CV APPLICATION COMPARISON

CREATIVE BOX GMBH

COMPARISON	CANDIDATE A: MARKETING MANAGER	CANDIDATE B: ART DIRECTOR	CANDIDATE C: PROJECT MANAGER
YEARS OF EXPERIENCE	3	8	5
PROFESSIONS	MARKETING MANAGER MARKETING SPECIALIST	SOCIAL MEDIA MANAGER DIGITAL MARKETING MANAGER	SEO/SEM SPECIALISTS PROJECT MANAGER
NUMBER OF PREVIOUS EMPLOYERS WORKING AT	2	3	2
LANGUAGES	ENGLISH, SPANISH, FRENCH	ENGLISH, FRENCH	ENGLISH, GERMAN
EDUCATION	B. A. BUSINESS ADMINISTRATION, M. SC. DIGITAL MARKETING	B. A. DESIGN M. A. MARKETING	B. SC. MARKETING, M. SC. DATA MANAGEMENT
ADDITIONALL SKILLS & ACHIEVEMENTS	CREATIVITY, COMMUNICATION, TEAMWORK, CRITICAL THINKING, RESILIENCES	SOCIAL AD CAMPAIGNS, WEB DESIGN, GRAPHIC DESIGN, SEO/SEM	EMPLOYEE OF THE YEAR, DATA ANALYSIS, LEADERSHIP, COMMUNICATION

You can consider the guidance offered by an AI agent in your HR department with expertise in recruitment decisions. The AI agent has conducted additional online research, including social media analysis, and compared the information on the candidates' CVs to predict the accuracy of specified competencies and assess the cultural fit of each candidate. Considering the predictions, the AI agent recommends hiring candidate C.

Figure 2. Example of Decision-Making Tasks

Participants

Data collection took place in January and February 2024, resulting in 215 leaders completing the survey with valid answers. To focus on leaders is perceived as providing insights into the group of individuals who face decision-making daily. This allows us to receive insights into responsibility attribution behavior and the impact of moral identity on such within a group that actually bears a certain responsibility for other people's lives, making the focus on leaders particularly interesting.

We recruited leaders via LinkedIn by reaching out to them through private messaging and sharing the survey link across different LinkedIn groups to promote participation. In addition, we recruited leaders on Prolific by using the filter options stating that they need to have a management position, speak English, and have decision-making power in their daily lives. This study focused on leaders as participants to explore their decision-making in organizational contexts, emphasizing the importance of their experience with real-world complexities. Leaders were recruited through both, Prolific and LinkedIn to ensure diversity, targeting leaders from various levels. Participants on Prolific received an average payment of $15.38 per hour for completing the survey. In total, valid responses were 71.16% from Prolific and 28.84% from LinkedIn. Participants were predominantly male (61%), aged 30–50, and held at least a bachelor's degree.

Measures

To measure *moral identity* as our independent (H2) and moderating (H3) variables, we relied on the scale suggested by Black and Reynolds (2016). According to Black and Reynolds (2016), moral identity can be understood as a combination of a person's moral self and the person's moral integrity, which the authors suggest measuring separately in their proposed moral identity scale, which we asked the participants to complete after completing all decision-making tasks. To measure *self-attribution of responsibility* as the dependent variable, the study relies on the responsibility attribution scale developed by Passlack et al. (2024). Their scale was adapted from the original responsibility attribution scale introduced by Hinds et al. (2004) and employs a 7-point Likert scale, ranging from 1 (indicating low self-attribution) to 7 (indicating high self-attribution) of responsibility, tailored to the context of decision-making; details on the measurement items can be found under OSF (https://osf.io/uh4rw). The *decision-making outcome* as a

dependent variable is a dichotomous variable for 0=success and 1=failure. In the experiment, this was a randomized treatment.

Data Analysis

To examine our hypotheses, we conducted a series of statistical analyses using R Studio, and significance levels were set at $p<0.05$ for hypothesis testing. The dataset comprised 108 participants who completed four tasks while receiving algorithmic advice and was analyzed using linear regression. For the first two hypotheses, simple linear regression was applied; for the moderating influence of the third hypothesis, an interaction term was included in the model. We tested for normality of residuals using the Shapiro-Wilk test with $W=0.95975$ and $p<0.001$, indicating that residuals are not normally distributed. This may suggest violations of the normality assumption of linear regression, which could affect the reliability of the model's inferences. Reliability of the variables has been tested using Cronbach's Alpha α, with results of all variables passing the 0.80 level (Cho & Kim, 2015).

As there had been four different tasks within the vignette experiment of Passlack et al. (2024), upon which this study relies, we further tested if there are differences in responsibility attribution across tasks. The mean values of responsibility attribution do not vary much across the four tasks, with values ranging between 5.21 and 5.41. Also, the standard deviations (SD) are similar across tasks, meaning the variation in responsibility attribution is consistent. We further checked differences between tasks using one-way ANOVA (Dennis & Valacich, 2001) with the following results: $F(3, 428)=0.813$, $p=0.487$. The p-value (0.487) is not significant, meaning there is no significant difference in responsibility attribution between the four tasks. This is why we ran our hypotheses validation across all tasks (n=432) and did not integrate the task type within the regression analysis.

Results

Table 1 provides an overview of the empirical results and our hypotheses validation.

Hypothesis	Predictor Variable(s)	Estimate	Std. Error	t-value	p-value	Significance	Conclusion
Regression	(Intercept)	4.002	0.636	6.296	< 0.001	***	-
H1	Outcome: failure (ref. success)	-0.307	0.106	-2.888	0.00408	p < 0.01	Supported
H2	Moral Identity	0.319	0.105	3.037	0.00253	p < 0.01	Supported
H3	Outcome*Moral Identity	-0.056	0.209	-0.270	0.7869	Not Significant	Not Supported

Table 1. Hypotheses Validation

With $R^2=0.0040$, the model explains little variance in responsibility attribution. With F-statistics=6.001 ($p=0.0005$), the overall model is statistically significant.

As can be seen in Table 1, the decision-making outcome as a predictor variable has a statistically significant effect on leaders' self-attribution of responsibility, supporting our first hypothesis (H1). Thereby, the results indicate that leaders in the failure condition attribute less responsibility to themselves compared to those in the success condition. The effect size ($R^2 = 0.019$) is small, meaning the decision-making outcome explains about only 1.9% of the variance in responsibility attribution. Regarding H2, the results show that leaders' moral identity significantly predicts their self-attribution of responsibility. The positive estimate (0.319) suggests that leaders with higher moral identity attribute more responsibility to themselves despite the decision outcome. The effect size ($R^2 = 0.021$) is small but slightly higher compared to the decision outcome, indicating moral identity explains about 2.1% of the variance in responsibility attribution. Concerning H3, which examined the moderating influence of leaders' moral identity on the relationship between decision outcome and responsibility attribution, the results revealed non-significant values ($p=0.787$). This indicates that moral identity does not moderate the relationship between decision outcomes and self-attribution of responsibility significantly.

To summarize, leaders attribute less responsibility to themselves in failure conditions than in success conditions (H1). Additionally, leaders with higher moral identity attribute more responsibility to

themselves in general (H2), but this does not moderate the relationship between outcomes and self-attribution of responsibility (H3). This means that moral identity has a direct effect on self-attribution of responsibility, but it does not change the way people react to success or failure in terms of self-attribution significantly.

Discussion and Contributions

This study contributes to the existing IS literature on self-attribution of responsibility in algorithmic decision-making by demonstrating how moral identity can shape self-attribution of responsibility among leaders. While prior research has outlined the asymmetry in responsibility attribution between success and failure cases (e.g., Mayer et al., 2020; Passlack et al., 2024), our findings extend this framework by integrating moral identity as a key determinant in reducing responsibility diffusions, highlighting the relevance of integrating moral identity theory with responsibility attribution theory. This study further adds to the field of human-AI interaction by addressing responsibility gaps emerging when individuals rely on algorithmic advice. Previous research outlined possible responsibility gaps (e.g., Dong & Bocian, 2024, 2024), and our findings contribute to this by suggesting that leaders with a strong moral identity are found to remain more committed to taking responsibility for their decisions, regardless of decision outcomes. This supports prior research emphasizing the need for integrating considerations of moral identity in human-AI interaction conceptualizations (e.g., Constantinescu et al., 2021; Heyder et al., 2023). Thereby, this study answers the call of McGuire and De Cremer (2023) for more examinations of what shapes leaders' perceptions of responsibility when engaging with algorithmic advice.

While we observed a direct effect of moral identity on the assumption of responsibility for outcomes, we found it interesting and important to highlight that the moderating effect of moral identity was non-significant. This may be attributed to the relatively homogeneous nature of our sample, which focused exclusively on leaders with similarly high levels of moral identity, potentially limiting the variance needed to detect significant moderating effects. Furthermore, participants' moral identity may not have been sufficiently activated in the experimental context, particularly if the task was not perceived as morally salient (De Freitas & Johnson, 2018). Future research should consider more diverse samples and decision-making scenarios where moral considerations are more explicit to better understand the conditions under which moral identity may be a moderating factor, having in mind that we found moral identity to have a direct effect on responsibility attribution.

In terms of practical implications, this study's empirical findings emphasize the importance of nurturing moral identity among decision-makers in organizations. As organizations increasingly integrate algorithmic support in decision-making processes, there is a risk that leaders may feel less accountable for outcomes over time, especially in cases of failure (Michael et al., 2024; Passlack et al., 2024). Our research suggests that moral identity can counteract this tendency, ensuring that human decision-makers take responsibility for decision outcomes. Hence, organizations may consider training programs that activate ethical self-concept, fostering a culture of shared responsibility (Cañas, 2022), and encourage decision-makers to reflect on their moral values even when using algorithmic guidance. Additionally, including moral identity assessments within leadership recruitment could cultivate ethical leadership.

Conclusion

This study provides empirical evidence on the influence of leaders' moral identity on self-attribution of responsibility when relying on algorithmic advice in their decision-making processes. While confirming the tendency of leaders to attribute more responsibility to themselves in success cases than in failures, it also demonstrates that leaders' moral identity is positively associated with self-attribution of responsibility, regardless of decision outcomes.

Besides, this study comes up with some limitations. First, the experimental vignette design may not fully capture the complexities of real-world decision-making. Future research could use field studies to validate these findings in different organizational settings. Second, the sample, though diverse in leadership roles, may not fully generalize across industries and cultural settings with different moral foundations. Expanding the scope to different domains could yield further insights. Third, the sample also demonstrated a high degree of homogeneity due to the specific focus on leaders, who were predominantly male and aged between 30 and 49. This limited diversity may have resulted in low variance in moral identity levels, which could

have contributed to the absence of a significant moderating effect. Future research could address this limitation by accounting for potential moderating influences of demographic variables and including more diverse participant groups. Additionally, examining a wider range of decision-making contexts may help to better understand the potential moderating role of moral identity.

Overall, this research contributes to IS research by deepening the understanding of responsible decision-making in environments where algorithmic advice is used. These insights encourage the integration of moral identity into theoretical frameworks on responsibility attribution and suggest that leaders' moral identity should be taken into account in practice when addressing the assumption of responsibility. As algorithmic advice becomes prevalent, ensuring that leaders uphold moral responsibility remains a critical priority.

References

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, *6*(3), 128–138.

Barnes, S. B. (1999). *Understanding Agency: Social Theory and Responsible Action*. 1–176.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3).

Black, J. E., & Reynolds, W. M. (2016). Development, reliability, and validity of the Moral Identity Questionnaire. *Personality and Individual Differences*, *97*, 120–129.

Bleher, H., & Braun, M. (2022). Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems. *Ai and Ethics*, *2*(4), 747–761.

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology*, *13*, 836650.

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha. *Organizational Research Methods*, *18*(2), 207–230.

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*.

De Freitas, J., & Johnson, S. G. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology*, *79*, 149–163.

Dennis, A. R., & Valacich, J. S. (2001). Conducting Experimental Research in Information Systems. *Communications of the Association for Information Systems*, *7*(5), 1–41.

Dong, M., & Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. *Journal of Experimental Social Psychology*, *111*, 104584.

Frimer, J. A., & Walker, L. J. (2009). Reconciling the self and morality: An empirical model of moral centrality development. *Developmental Psychology*, *45*(6), 1669.

Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, *36*(1), 2009222.

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), Article 4.

Greene, P., Latham, A. J., Miller, K., & Norton, J. (2022). Capacity for simulation and mitigation drives hedonic and non-hedonic time biases. *Philosophical Psychology*, *35*(2), 226–252.

Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems*, *35*(4), 1092–1120.

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*.

Hardy, S. A., & Carlo, G. (2011). Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action?: Moral Identity Development. *Child Development Perspectives*, *5*(3), 212–218.

Harvey, J. H., Weary, G., & Harris, B. (1981). *Perspectives on attributional processes*. W.c. Brown Co.

Heider, F. (1958). *The psychology of interpersonal relations* (pp. ix, 326). John Wiley & Sons Inc.

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.

Hinds, P., Roberts, T., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, *19*(1), 151–181.

Horneber, D., & Laumer, S. (2023). Algorithmic Accountability. *Business & Information Systems Engineering*, *65*(6), 723–730.

Kegel, M., & Ghanem, L. (2024). *You Did That on Purpose! An Investigation of the Knobe Effect in Human-Robot Interactions*.

Lapsley, D. K., & Narvaez, D. (2004). A Social–Cognitive Approach to the Moral Personality. In *Moral Development, Self, and Identity*. Psychology Press.

Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, *19*(4).

McGuire, J., & De Cremer, D. (2023). Algorithms, leadership, and morality: Why a mere human effect drives the preference for human over algorithmic leadership. *AI and Ethics*, *3*(2), 601–618.

Michael, K., Schoenherr, J. R., & Vogel, K. M. (2024). Failures in the Loop: Human Leadership in AI-Based Decision-Making. *IEEE Transactions on Technology and Society*, *5*(1), 2–13. IEEE Transactions on Technology and Society.

Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, *23*(4), 579–587.

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.

Nicholls, J. G. (1975). Causal attributions and other achievement-related cognitions: Effects of task outcome, attainment value, and sex. *Journal of Personality and Social Psychology*, *31*(3), 379.

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*, *39*(4), 1871–1882.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors The Journal of the Human Factors and Ergonomics Society*, *64*, 1–35.

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.

Passlack, N., Heyder, T., Klemm, F., & Posegga, O. (2023). How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success. *ICIS 2023 Proceedings*.

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.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28.

Rowe, M., Wilcox, L., & Gadlin, H. (2009). *Dealing with—Or Reporting—"Unacceptable" Behavior*.

Schwartz, M. S. (2016). Ethical Decision-Making Theory: An Integrated Approach. *Journal of Business Ethics*, *139*(4), 755–776.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411.

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), Article 2.

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*, *22*(2).

Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9.

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.

van der Vlist, F. N. (2016). Accounting for the social: Investigating commensuration and Big Data practices at Facebook. *Big Data & Society*, *3*(1), 2053951716631365.

Weiner, B. (1985a). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573.

Weiner, B. (1985b). "Spontaneous" causal thinking. *Psychological Bulletin*, *97*(1), 74–84.

Paper V

Passlack, N. (Working Paper, unpublished). From Teachers to Cheerleaders? Teachers' Role Responsibilities in the Age of Generative AI.

From Teachers to Cheerleaders? Teachers' Role Responsibilities in the Age of Generative AI

Completed Research Paper

Abstract

GAI tools are opening up new opportunities in education, reshaping how learning takes place and how teaching is practiced. This study examines teachers' experiences with GAI, exploring how its use transforms specific areas of teaching and influences how teachers perceive and redefine their professional role responsibilities in the age of GAI. Grounded in a sociomateriality perspective, the paper adopts a relational view, focusing on the dynamic interactions between teachers and GAI. Based on interviews with 28 secondary teachers, the findings reveal four distinct ways in which teachers make use of and perceive GAI: (i) as a supportive tool, (ii) as a critical friend, (iii) as a disruptive presence, and (iv) as a sparring partner. These perceptions are closely linked to four identified corresponding professional role responsibilities: (i) consulting, (ii) cheerleading, (iii) enlightening, and (iv) challenging. In doing so, the study contributes to IS research by showing how GAI may reshape perceived role responsibilities and work processes among secondary teachers, who are responsible for shaping our future, the professionals of tomorrow. This results in four propositions for future research.

1 Introduction

With Generative Artificial Intelligence (GAI) entering the daily realities of educational practices, it not only requires teachers to adapt to new tools but also to renegotiate what it means to teach. GAI, drawing on technologies such as natural language processing, neural networks, and deep learning (Benbya et al., 2021), enables teachers to conduct rapid research, develop teaching proficiency (Jaiswal & Arun, 2021), reflect on their decisions (Aldeman et al., 2021), or facilitate brainstorming for more creative tasks. This dynamic is expected to alter teacher-student interaction (Gentile et al., 2023) and comes with certain ethical implications (Lameras & Arnab, 2021). For instance, research has shown that technology-mediated decision-making can alter professionals' sense of agency and relevance, leading professionals to feel less useful in their roles (Strich et al., 2021). Consequently, if not adequately understood and addressed, the use of GAI may erode teachers' self-satisfaction and diminish their educational commitment.

While an increasing body of valuable research has examined the potential of AI in education (Adiguzel et al., 2023; BaiDoo-Anu & Owusu Ansah, 2023; Yusuf et al., 2024), its implications for teachers' professional role responsibilities and the resulting consequences remain largely unexplored. The increasing delegation of tasks to GAI, coupled with its opacity, can create uncertainties regarding accountability for

outcomes (Giermindl et al., 2022) and a shifting sense of professional identity and role responsibilities among teachers. This tension seems particularly salient in vocational education, where teachers not only impart technical competencies but also serve as mentors and role models for future professionals. Thus, with increased GAI use, we see secondary teachers find themselves navigating complex technological, pedagogical, and ethical terrains. It is therefore essential that we better understand how teacher re-craft their teaching by embedding GAI technologies into their teaching practices and how this affects their perceived professional role responsibilities, since this has effects on the meaningfulness of their work (Harder Fischer et al., 2024). By professional role responsibilities, this paper refers to the duties and responsibilities associated with an individual's professional role, which are shaped by commonly held expectations about how someone in a specific position should behave (Stets et al., 2020).

Sparked by increased capabilities of GAI models like ChatGPT 5, raising provocative questions like, "*Why do we still need teachers if we can turn to ChatGPT for every answer?*", this study seeks to reimagine and redefine what it means to be a teacher in the age of GAI, questioning GAI's authority. While acknowledging that GAI can substitute certain aspects of teachers' work, this study argues that such substitution occurs meaningfully only through collaboration between teachers and GAI. In doing so, the study follows a relational view of GAI, meaning that the capabilities of the technology and its users evolve through collaboration (Orlikowski, 1992). This implies that AI's capabilities are not fixed (Heyder et al., 2023; Hinds & Von Krogh, 2024) but instead develop through teachers' practical use and perceptions. Rather than speculating from the outside, this study argues that we need to turn to those at the heart of the profession, teachers who are already using GAI technologies in their daily practice. By listening to their experiences, this study aims to explore not only how teachers (re)craft their working practices with GAI, but also how they see their professional role responsibilities changing in this new educational landscape. Specifically, this study addresses the following research questions:

RQ: *How do teachers perceive GAI reshaping their working practices and professional role responsibilities?*

To answer the research question, the study draws on 28 interviews with secondary teachers to examine how the nature of educational work is changing through the use of GAI and to theorize on the potential long-term implications for the evolution of teachers' professional role responsibilities. In doing so, the study seeks to explore the "digital futures" of secondary teachers, as they are the ones who shape the future of so many. Recognizing that the use of GAI in educational practice remains in its early stages, this work represents an initial inquiry into how teachers envision the transformation of their professional role responsibilities, adapting a rather speculative 'future orientation' (Hovorka & Mueller, 2025). In line with Schlagwein et al. (2025), this study adopts the notion of futures in the plural, acknowledging the inherent

uncertainty in predicting how human-GAI collaboration may transform social subsystems. In doing so, this study envisions diverse digital futures, shaped by teachers' mental models of perceived shifts in their professional role responsibilities.

The study seeks to contribute to the existing body of Information Systems (IS) research by learning from practice to shed light on how we need to redefine professional role responsibilities of secondary teachers due to the immediate societal multiplier effect of teachers GAI use as shifts in teaching, especially in how the next generation of students learns to work with and think about GAI. The findings identify four ways in which teachers make use of and perceive GAI (as a supportive tool, critical friend, disruptive presence, and sparring partner), associated with distinct types of role responsibility (consulting, cheerleading, enlightening, and challenging). These links reveal how GAI use can influence the ways secondary teachers construe and perform their professional role responsibilities. In doing so, it offers propositions for how different patterns of GAI use may shape responsibility perceptions, providing a foundation for future IS research on human-GAI collaboration and the evolution of professional roles in technology-rich environments. In doing so, the study informs strategies for supporting teachers in navigating the ethical and identity-related implications of GAI integration.

The remainder of the paper is organized as follows: it begins with a concise overview of the theoretical background (section 2), followed by a description of the methods employed (section 3). The subsequent section presents the findings (section 4), which are then examined in the discussion, highlighting the key contributions of this work (section 5). The paper concludes with a brief summary (section 6).

2 Theoretical Background

To explore teachers' digital futures with respect to their professional role responsibilities in the age of GAI, the following sections present the theoretical background necessary to examine the phenomenon. Section 2.1 introduces the theoretical lens through which digital futures are conceptualized, while section 2.2 outlines the key constructs relevant to this study.

2.1 Understanding teacher-GAI collaborations as sociomaterial practice

This study takes a relational view on human-GAI collaborations, meaning that GAI's capabilities are not static but evolve through how teachers use and interpret them in practice (Heyder et al., 2023; Hinds & Von Krogh, 2024). This goes in line with research that has emphasized that "AI [should not be abstracted] out of the context of its existence and use, ignoring the human beings and the social institutions and social arrangements that give AI its capacity to do or mean anything," which referred to "sociotechnical blindness" (Johnson & Verdicchio, 2017, p. 2268). In doing so, this study conceptualizes teachers' collaboration with

GAI as part of a socio-technical system, in which teachers (as social entities) interact with GAI-based tools (as material entities) referred to as sociomaterial practice (Leonardi, 2012). This sociomaterial practice, where the social (e.g., teachers, practices, responsibilities) and the material (e.g., GAI technologies, tools) are inseparably intertwined in shaping work and meaning, is often referred to as sociomaterial entanglement or "imbrication" (Leonardi, 2012). Sociomaterial entanglement is expected to shape the broader social subsystems (Leonardi, 2012), such as educational organizations, including values, roles, processes, and role responsibilities.

Following the work of Jones (2014) and Heyder et al. (2023), and adapting it to the context of teachers, this paper posits that teachers using GAI engage in a sociomaterial entanglement, characterized by a state of intertwining and *inseparability* between teachers' and technological agency. The integration of GAI into education is expected to influence teachers' practices and routines, which is linked to the notion of *practices*. Teachers' perceptions of GAI are shaped by the societal subsystems that influence their interactions with the technology, reflecting the concept of *performativity*, how humans construct and enact their reality. These perceptions, in turn, are expected to influence the nature of teachers' engagement with GAI, further reinforcing their *inseparability*. Finally, teacher-GAI interactions can affect teachers' sense of agency and contribute to potential shifts in their professional role responsibilities, a dynamic described through the notion of *relationality*.

Concerning the sociomaterial entanglement, this study differentiates between automation (AI overtakes complete tasks of teachers, e.g., automated exam evaluation by multiple-choice tests) and augmentation (AI enhances tasks of teachers, e.g., supporting the development of learning concepts). Although recognizing that the two are inherently interdependent and cannot be fully separated (Raisch & Krakowski, 2021). Following calls for a process-oriented perspective on human-GAI interaction (Heyder et al., 2023), the distinction is nonetheless useful here as it helps to capture teachers' perceptions. It reflects the extent to which teachers perceive themselves as actively engaged in completing a task versus viewing the GAI as executing the task on their behalf, an orientation expected to shape both GAI perceptions and perceived role responsibilities. In this context, *automation* describes situations in which teachers delegate a task to GAI that they could have completed themselves with comparable results. By contrast, *augmentation* refers to instances where GAI is seen as enhancing human capabilities, with teachers recognizing that the GAI's output surpasses what they could have produced unaided, particularly in terms of quality (Rai et al., 2019).

2.2 On Job Crafting Affecting Professional Role Responsibilities

As GAI becomes increasingly integrated into the field of education, it inevitably compels teachers to re-craft their working routines. GAI integration is becoming increasingly democratized (Krakowski, 2025),

meaning it can be initiated by teachers themselves and implemented in the classroom through top-down directives, but also emerges from the bottom up through student adoption, which leads to continuous changes and requires ongoing readjustment.

Job crafting is a concept that describes how professionals actively shape and redefine their work (Tims et al., 2012). It is a bottom-up approach as it refers to "self-initiated change behaviors that employees engage in with the aim to align their jobs with their own preferences, motives, and passions" (Tims et al., 2012, p. 173). While jobs are associated with specific duties, such as teachers being responsible for delivering particular learning content, there is often flexibility in how these duties are fulfilled. The age of GAI presents significant opportunities for teachers to streamline certain work processes (Hu, 2021), adopt more innovative approaches (Chen et al., 2023), and redirect their efforts toward new tasks. These developments can be understood through the lens of job crafting. Concerning this, "teachers' perceptions, acceptance, knowledge, and practices regarding GenAI are interrelated factors that play central roles in teachers' uses of GenAI" (Zhai, 2024, p. 5). Job crafting further is expected to have consequences on how professionals identify with their jobs in the sense of meaningfulness (Harder Fischer et al., 2024), thereby affecting their professional role identity.

Role identity theory, instead, explains that individuals construct a sense of self through the enactment of these roles, which provide meaning and purpose in life (Strich et al., 2021) and signal to both the self and others how one fits into society. Professional role identity is defined in answering the questions 'Who am I (as a member of a specific profession)?' and 'What do I do?'" (Strich et al., 2021, p. 304). In other words, it defines the aspect of a person's self-concept that is derived from the social roles they occupy and the expectations associated with those roles (Stets et al., 2020). It includes the internal meanings, attitudes, values, beliefs, behavioral norms, and responsibilities that individuals associate with fulfilling particular roles, such as being a teacher.

Responsibilities associated with one's job constitute an integral part of one's professional role identity. As summarized and outlined by Cañas (2022), responsibility generally refers to a person's duty or obligation to carry out a task. It can be defined by their role (e.g., as part of a job), often grounded in ethical norms, and it is not always formalized in rules or regulations. Accountability, while sometimes used interchangeably with responsibility, is different in that it involves answering to others. It is tied to external expectations or authority, meaning a person must explain and justify their actions to someone in a position of oversight. Liability is a specific type of responsibility with a legal dimension. It is defined by laws, regulations, or contractual agreements and is enforceable through formal legal processes.

By focusing on how teachers perceive changes in their role responsibilities in the age of generative AI (GAI), this study examines responsibilities that are not legally binding but are perceived by teachers as

socially expected or necessary to ensure responsible education. Accordingly, this paper uses the term professional role responsibilities (as part of an individual's role identity) to refer to the duties associated with a teacher's professional role. The paper assumes that perceived professional role responsibilities influence the way teachers craft their jobs, while job crafting, in turn, shapes how teachers perceive their role responsibilities.

3 Methods

This study relies on semi-structured interviews with 28 secondary teachers for several reasons. Concerning the focus on secondary teachers is perceived as critical, as they are the ones who are directly responsible for educating and preparing the next generation of skilled workers. Their professional role identities influence student learning, and their GAI use also affects how future professionals understand and use GAI in their own fields. Further, teachers in general can be perceived as frontline ethical agents as they make daily decisions about how GAI is used in classrooms by being confronted with questions like 'What content is appropriate?' and 'How do we ensure fairness and inclusion?'

The decision to employ semi-structured interviews for addressing the research questions is based on two main considerations. First, professional role identities represent a highly complex phenomenon that remains insufficiently understood; therefore, qualitative research methods are essential for generating theory and developing preliminary propositions (Sarker et al., 2013). Second, professional role responsibilities are closely tied to subjective perceptions. Consequently, semi-structured interviews are appropriate to allow deeper exploration and provide researchers the flexibility to probe further when individual responses necessitate clarification or additional elaboration (Sarker et al., 2013).

The following sections outline the interview guideline (section 3.1), details on the participants and sampling (section 3.2), and the data analysis by outlining the coding approach (section 3.3).

3.1 Semi-Structured Interview Guideline

A semi-structured interview guide was designed to explore teachers' perceptions, practices, and ethical considerations regarding the use of GAI in education. The semi-structured interview protocol helped to maintain consistency while allowing for in-depth exploration (Myers & Newman, 2007). It combined open-ended questions with conceptual anchors to ensure both flexibility and theoretical relevance. The guide was structured into thematic sections, beginning with demographic and contextual information about the participants' professional background, teaching experience, and subject areas. This was followed by questions aimed at assessing participants' understanding of GAI applications and identifying specific areas of use in their daily teaching practice. Core sections of the interview addressed perceived changes in teachers' professional roles and responsibilities in response to GAI integration, with particular emphasis on

the redistribution of responsibility between humans and GAI systems. Additional items examined participants' ethical concerns, as well as how these concerns influenced their classroom practices. The final section focused on identifying necessary support structures and capacity-building measures to promote responsible GAI use in education. The complete interview guide is provided in Appendix I. The interview guideline was tested with two Pilot interviews to guarantee understandability.

3.2 Sampling and Participants

Data collection took place in January and March 2025, resulting in 28 teachers participating. Secondary teachers were recruited by approaching school rectors through emails and, in short, explaining what the study was about, thereby following a convenience sampling strategy. All interviews were conducted in German and translated into English afterwards.

The focus on secondary teachers is considered essential, as they are responsible for preparing the professionals of tomorrow. Equipping individuals as effectively as possible is seen as a key factor in shaping a competent workforce for the future economy. This study, therefore, explores how secondary teachers can harness the potential of GAI in education and how their responsibilities evolve in the context of responsible teaching. Secondary teachers are viewed as particularly important in this regard, as they can articulate the ethical concerns they face, and reflect on how their professional role identities may (or should) change to ensure that GAI is used responsibly in educational settings.

The interviews were held in German via video conference in Teams or Zoom, and quotes were translated into English afterwards. All teachers had a minimum of five years of work experience, regularly used GAI tools for teaching, and were between 30 and 62 years old. Appendix II provides an overview of the demographics.

3.3 Data Analysis

This study's data analysis adopts an exploratory approach grounded in constructivist theory, emphasizing an interpretivist perspective, following the definitions by Schlagwein et al. (2025). It aims to outline possible futures for teachers by analyzing their narratives regarding various teaching situations with GAI, while closely adhering to the teachers' original wording. The questions this study seeks to answer concern the future of education by analysing "Where are we?" and "Where are we heading?" (Raskin et al., 2002).

The transcribed interviews were coded according to what Gioia (2013) frames as Open-Axial and Selective Coding. This approach seems suitable since the answers were very subjective and individual, as the study has a focus on theorizing changes in professional role responsibilities by considering teachers' perceptions and narratives. The approach allows for remaining very close to the framing of the participants and to mirror personal subjective views (Gioia, 2013, 2021). It allows allows to make sense of the data by identifying

and elaborating on the diverse perspectives regarding how GAI might reshape teachers' professional role identities.

In the *initial phase* of the analysis, the empirical data were carefully read, and open coding was applied to identify different GAI use cases and role responsibilities. The finding of so-called first-order concepts was grounded in the participants' language, maintaining a close alignment with how teachers described their experiences. This informant-centered strategy aimed to reduce researcher bias and ensure that the insights were deeply rooted in the respondents' perspectives (D. Gioia, 2021). The *second phase* shifted toward organizing and interpreting the first-order codes into broader second-order themes. This stage involved identifying conceptual patterns that could offer a more theoretical understanding of participants' experiences (D. Gioia, 2021). While the first phase emphasized descriptive insights, this second step focused on analytical interpretation, clustering the data into categories that summarize the first-order concepts (D. A. Gioia et al., 2013). The *third phase* involved abstracting the themes into overarching dimensions, derived through an iterative analysis of the second-order themes. These dimensions reflect four distinct types of GAI perceptions in educational use, as well as four types of role-related responsibilities. Finally, in the *fourth phase*, mapping was carried out to link the identified role responsibility types to the GAI perceptions. An overview of the data structure can be found in Appendix III-IV.

4 Findings

The following sections present key findings from interviews conducted with 28 secondary teachers. To answer the research question, it is important to recognize that the implications of GAI on professional role responsibilities largely depend on how it is used and perceived (as outlined in Section 2.1), and these factors are inherently individual. Thus, before exploring how teachers perceive shifts in their professional role responsibilities, Section 4.1 outlines the areas in which GAI is currently applied in vocational education and assesses the extent to which these applications impact teaching tasks, specifically, whether GAI augments teachers' capabilities by enabling innovative teaching formats or primarily automates tasks. Then, section 4.2 examines how teachers perceive changes in their professional role responsibilities.

4.1 Emerging teaching task formats through teachers' job crafting: Four types of GAI perceptions

The findings revealed a variety of tasks for which secondary teachers make use of GAI, and Table 1 provides an overview. To gain deeper insight into how teachers' job crafting through GAI may reshape their professional role responsibilities, the tasks involving GAI were categorized along two key dimensions: (i) the nature of human-GAI collaborations, differentiating between automation and augmentation, as defined in section 2.1., and (i) the innovativeness of the GAI application, distinguishing between familiar and emerging teaching tasks. *Familiar tasks* refer to activities that teachers were already performing before

integrating GAI into their practice, and *emerging tasks* represent new and innovative areas of activity that have become possible or relevant only through the use of GAI. Table 2 provides an overview of the findings, which are explained in detail in the following sections.

Table 1. Four clusters of GAI applications linked to GAI perceptions

	Familiar teaching tasks	***Emerging teaching tasks***
Automation	**GAI as supportive tool** • Structuring lesson plans • Producing informational materials • Creating assessments and exams • Developing practice exercises • Provide learner feedback	**GAI as disruptive presence** • Visualizing content through GAI-generated images • Producing educational podcasts • Producing learning videos
Augmentation	**GAI as critical friend** • Searching and retrieving instructional content • Translating learning materials • Delivering differentiated instructions	**GAI as sparring partner** • Co-creating ideas with GAI as thinking partner

GAI as a supportive tool

As the data indicates, the majority of teachers use GAI primarily as a supportive tool to automate tasks (*automation*) they would typically perform manually (*familiar teaching tasks*). GAI as a supportive tool implies that the GAI systems are used to assist with tasks without replacing teachers' autonomy. This use is largely driven by a desire "*to save time*" (T6) and to redirect their efforts toward more complex or pedagogically valuable activities ("*it makes my work easier and allows me to use the time more intensively with the students*" (T7)). "*I'm happy to let it handle the really annoying tasks.*" (T21), as a teacher framed it. The following section outlines the five main tasks for which AI was used as a supportive tool by the teachers interviewed (generating learner feedback, developing practice exercises, creating assessments and exams, producing informational materials, and structuring lesson plans).

Tasks where teachers use GAI as supportive tools were, for instance, **structuring lesson plans**: "*If I ask the GAI to make a suggestion on how to structure a unit on topic XY, of course it's much quicker than spending half an hour racking my own brain before I have even a rough plan. Then I can take what the GAI suggests and modify it to fit my specific student group and the content I want to include*." (T15) Also, GAI tools such as ChatGPT help some teachers to **produce information materials** where teachers use the help of GAI: "*For example, I often give it my own materials and have it rephrase them nicely or change the*

format, so that, for instance, an interview is created, which is sometimes more interesting for the students to read than plain text." (T6)

Beyond producing informational materials, teachers also experiment with **creating assessments and exams** using GAI. As one teacher shared, "*I once had a performance test created as a trial*" (T19). However, generating high-quality exams with GAI remains challenging. Another teacher noted, "*I've also tried exams before, but so far I've struggled a bit with that. They still don't meet the qualitative standards*" (T9). In contrast, GAI proves to be more effective for developing case studies or dialogue-based assessment materials. One example would be "*Tasks I've had saved for years, which might still be relevant, I just put into the AI and say, 'Create an exercise in this style,' and then it's simply adapted.*" (T25).

Several teachers reported using GAI tools to **develop practice exercises**. Examples included "*a quickly created quiz that also enables individualized learning*" (T9) and "*multiple-choice questions that can then also be used in Kahoot*." (T13). One teacher emphasized GAI's capacity to extract transcripts from videos and subsequently create exercises, describing this as "*something that used to be incredibly time-consuming*" (T18). Another one described an approach on how to generate a bingo game: "*I had thought that I would just do a bingo game with them, and I made it easy for myself by asking the AI: 'Please give me 12 more rhetorical devices.' I had specified beforehand which ones I wanted, like 'make me a metaphor for the rhetorical device,' 'make me a climax rhetorical device,' and 'a personification.' The AI did that, and then I asked it: 'Can you please generate a text for me using these rhetorical devices?' That's what I did. And so, within an hour, the students could play bingo and had to identify the rhetorical devices.*" (T17)

A smaller group of teachers **provided learner feedback** "*when it comes to assessing learning products, for example, to have it [GAI] generate feedback text modules that fit my specific learning product*." (T15) Yet, critical voices are also raised, as one teacher described: "*I actually tried it once with exam grading, by entering the grading rubric. I found that it definitely revealed clear gaps, especially with students who are not yet able to express themselves well linguistically, but whose content was actually quite good. In these cases, the AI definitely couldn't keep up, as it placed much more emphasis on linguistic expression*." (T17)

GAI as a critical friend

In contrast, some participants described GAI not merely as a tool, but what can be frames as a "critical friend" that enhances their existing practices (in the sense of *augmentation*), particularly in tasks that are familiar within the teaching profession (*familiar tasks*). The following section outlines the three main tasks for which AI was used as a critical friend by the teachers interviewed (searching and retrieving instructional content, translating learning materials, and delivering differentiated instructions).

GAI as a critical friend is most often linked to GAI that makes it possible to "*get ideas*" (T5), to "*ask about certain things*" (T13), and GAI that facilitates "*obtaining information even more easily, as was previously the case with Google*." (T3), this means it is used for **searching and retrieving instructional content**. Compared to traditional search engines like Google, GAI was perceived as more convenient due to its ability to synthesize information from multiple sources and respond to follow-up questions, like a "critical friend." "*I had one or two new topics that I had never taught before. In the past, I would spend hours gathering all my schoolbooks and going through all the materials I had. Now it's much faster, because I often ask the AI when I have a certain topic*" (T9). Some teachers agree, "*If I need inspiration for a presentation, I might ask Gamma App*" (T7) thereby supporting teachers in refining their ideas or decisions. Other teachers search for feedback when grading exams, "*[after] encountering the same supposedly incorrect spelling for the 300th time, you start to wonder whether you might be the one spelling it incorrectly. In such cases, I also ask the AI: 'Please do some internet research, check if there is any rule, a spelling rule, that says the word is spelled in such and such a way*" (T15).

GAI is also often used for **translating learning materials** into English. "*For example, when I'm grading something, an English exam, and I'm unsure about something, I often ask the AI. It then shows me that, yes, you can write it that way or say it like that, because in slang it's correct in this or that context.*" (T6).

Other teachers use AI tools to **deliver differentiated instructions** in the sense of "*text optimization*" (T3), as one teacher noted, "*I believe it is primarily the aspect of differentiation that AI greatly simplifies.*" (T3). GAI assists in varying the level of difficulty in tasks, enabling teachers "*to design tasks that are sometimes easier, sometimes more difficult for different groups of students*" (T25) and to provide "*additional explanations for weaker students*" (T13). In practice, this often involves refining tasks through prompts such as: "*The task is fine, make it more understandable or simpler for me, or rephrase it in a simpler way*" (T25). This interactive capability allowed teachers to iteratively refine their inquiries, thereby deepening their engagement with the content.

GAI as a disruptive presence

A minority of teachers reported that GAI also makes them dive into new teaching tasks by using GAI to try out new things (in the sense of *emerging teaching tasks*) without much involvement (linked to *automation*). In doing so, they described GAI rather as a disruptive presence, which implies a transformative impact of GAI that fundamentally alters teachers' teaching style or pedagogical approach. The following section outlines the three main tasks for which AI was perceived as rather a disruptive present by the teachers interviewed (visualizing content through GAI-generated images, producing educational podcasts, and producing learning videos).

Some teachers use GAI for **visualizing content through AI-generated images**, thereby it extends teachers' capability, as one said: "*I'm really bad when it comes to anything artistic. I just don't have an eye for aesthetics or design. I was never good at art in school either. So, in that area, I definitely see AI as something that can support me and help produce something decent*." (T20).

Thereby, GAI opens up new areas of teaching practice where many teachers previously had little or no expertise, such as **producing podcasts** or **instructional videos** (T12). The integration of such tools has also diversified classroom formats. As one teacher explained, "*I now often give my students the option to listen to the information as a podcast instead of just reading a text. I know myself, that's something I would never have done before if it weren't for AI.*" (T6). Another shared their experience creating a podcast recently: "*I was really impressed by how easy it was and how precisely it could be tailored to my lesson. I liked that*" (T23). Finally, one teacher described their workflow: "*If I already have the idea and just give the AI my starting point, then I can get a script for a video written more quickly, maybe even right away.*" (T27). *"So, my lessons have definitely become a bit more varied because of it" (T6),* as another stated.

GAI as a sparring partner

Other teachers described using GAI for tasks they had not previously undertaken (*emerging teaching tasks*), which extend their capabilities in ways they would not have achieved without GAI, but in cocreating processes, it helps them to reach superior outcomes (*augmentation*). In doing so, they use GAI as a sparring partner in the sense of an interactive counterpart that stimulates reflection, challenges ideas, or supports decision-making by offering alternative perspectives, feedback, or simulations. The following section outlines the task for which AI was used as a sparring partner by the teachers interviewed (co-creating ideas with GAI as a thinking partner).

Several teachers emphasized the value of GAI chatbots like ChatGPT as a reliable and readily available exchange tool, helping in **co-creating ideas with AI as a thinking partner**. One teacher noted, "*You always have a conversation partner, someone to exchange ideas with, available in some form*" (T5), highlighting GAI's role in facilitating ongoing dialogue. Teachers reported actively engaging the GAI with prompts such as "*How would you do it?*" (T5), leading to responses they described as inspiring and practically helpful: *"Then I get input that's valuable to me, where I think, 'Oh, cool idea, I've never tried that before.' It's kind of a substitute for collegial exchange, which often isn't possible due to time constraints, but which also brings in different perspectives from the entire data pool, so to speak.*" (T5) Such exchanges were perceived as intellectually stimulating and thought-provoking: "*Such an exchange really helps to sharpen one's thinking and come up with new ideas*." (T5) In doing so, teachers described

GAI as offering a second opinion on their teaching practice: "*When I work with AI, it's a bit like always having a second perspective on my teaching, my materials, and the content*" (T6). This role also extends into emotional and professional reassurance. As one teacher with many years of experience shared: "*Especially when you've been in the profession for so many years, it's just nice to have someone supporting you*" (T21) These insights suggest that for some teachers, particularly more experienced ones who may feel hesitant to seek help from younger colleagues, GAI can serve not only as a practical assistant but also as a form of emotional support, a sparring partner, and even a non-judgmental companion in the teaching process.

4.2 Teachers' emerging professional role responsibilities: Four types of role identities

A closer look at the data outlines how teachers' perceptions of their professional role responsibilities depend on how they use GAI. To gain deeper insights into how teachers' job crafting may reshape their professional role responsibilities, the duties that teachers perceived as arising from GAI integration were categorized along the same two key dimensions: (i) the familiarity of the professional role responsibilities; and (ii) the nature of human-GAI collaborations, differentiating between automation and augmentation.

Upon analysis of the reported perceptions on role responsibilities, four possible futures of role identities were identified, where each has its focus. Table 3 serves as an overview of the key findings, which will be outlined in the following.

Table 2. Four clusters of role identities

	Familiar responsibilities (despite GAI use)	***Emerging responsibilities (GAI-induced)***
GAI as substitution (automation)	**Teachers as cheerleaders** • Motivates students to remain engaged and confident in their learning journey. • Promotes and maintains strong interpersonal relationships with students.	**Teachers as enlighteners** • Highlight GAI's threats such as filter bubbles and fake news online. • Promotes social values
GAI as support (augmentation)	**Teachers as consultants** • Acts as a learning consultant, guiding students through the learning process. • Serves as a role model for responsible digital behavior	**Teachers as challengers** • Promotes foundational GAI literacy and understanding. • Identifies areas where students can improve and grow. • Encourages critical thinking by prompting students to question GAI-generated outputs.

Teachers as Cheerleaders

Some narratives of the interviewed teachers emphasized that, in the age of GAI, teachers' role responsibilities increasingly revolve around providing emotional support, someone who offers orientation and motivation to keep going, and being a consistent presence for students. Perceived role responsibilities were linked to motivating students to remain engaged and confident in their learning journey and promoting and maintaining strong interpersonal relationships with students.

As one teacher understood the importance of **motivating students to remain engaged and confident in their learning journey** framed it: "*I think leaving that [providing feedback] solely to an AI would be a serious mistake. I'd prefer […] the kind of feedback we provide through human interaction, ideally face to face, just the two of us. I see that as really important, based on how things work in our context. For example, when students […] send me their assignments, there's a small field for feedback. Sometimes that feedback is simply motivational*" (T7). By agreeing on the importance of motivating students, one teacher describes it as a broader, general challenge in today's educational context, one that cannot be addressed or resolved through the use of GAI: "*Honestly, I've noticed that today's vocational school students tend to be less capable. In my experience, we're seeing a higher number of weaker candidates compared to the past, students who, in some cases, really need a lot of support and, let's say, even a bit of a push to adopt a proper work ethic. So, to say that AI is taking over that part of my job, I honestly don't see it that way.*" (T18).

In general, teachers perceive that GAI also enables them to focus more on **maintaining strong interpersonal relationships with their students**, which are now more important than ever: "*Now I simply have much more time for the personal needs of the students, because those personal needs are growing, they're carrying heavier and heavier backpacks, so to speak. And that just gives me more time to support them.*" (T22) "*and also to work more on building relationships with them.*" (T6) Yet, some critical voices see that through GAI it may become more difficult to hold interpersonal relationships, despite agreeing that those interpersonal relationships are critical, as the following quote demonstrates: "*I find that somewhat troubling, the idea that humanity gets lost, also in the student-teacher relationship, perhaps, or even among the students themselves. To be honest, I'm a bit concerned about that.*" (T17)

Teachers as Consultants

Several accounts from the interviewed teachers highlighted that, in the era of GAI, their role responsibilities are increasingly centered on acting as a learning consultant by guiding students through the learning process

and serving as a role model for responsible digital behavior. Some teachers perceive themselves as what they refer to as a "*learning coach*" (T10), or "*learning consultant*" (T7), or "*moderator or supporter*" (T3), exemplified by statements such as, "*Now I talk to students about learning how to learn. I didn't do that in the past, of course, not in that way, because I simply didn't have the capacity for it.*" (T7).

One teacher notes that, in the age of GAI, their role is no longer centered on providing information. Instead, students are becoming more independent, and teachers must engage with them on equal footing, focusing on **guiding students through the learning process** by consulting self-directed and empowered learning, by also stating: "*At our school, we're currently focusing on things like learning spaces and creating an environment where students are given more freedom. They're allowed to go into side rooms, listen to music with headphones, eat or drink something, make themselves comfortable while learning. I think it becomes clear that, no, we're not their buddies. We're learning coaches, but we do interact with them on equal footing.*" Another one agrees: "*I believe that our understanding of the teacher's role, standing in front of students, is already shifting toward becoming a learning facilitator, a manager, a support tool; call it what you will, but 'learning facilitator' is probably the best term. […] Students should generate much more on their own, so that they should be action-competent and demonstrate competencies rather than just subject knowledge. And that essentially implies that the teacher, as a mere transmitter of information, is becoming obsolete and instead stands as a learning facilitator.*" (T7) This also implies, "*that we are actually moving toward a much more student-centered form of teaching, and that my role as a teacher becomes more that of a moderator or supporter, rather than someone who has to centrally steer the lesson and try to offer something for everyone at the same time.*" (T3)

Therby, teachers see increased responsibilities linked to **serving as a role model for responsible digital behavior**. As one teacher expressed their concern: "*It doesn't help to just let them loose with it. They think, 'Oh, life is easy, I don't have to think for myself anymore.' I really think the responsibility of schools is enormous, because otherwise, we'll end up with a bunch of stupid, unreflective people*." (T25, Pos. 2). Likewise, one teacher expressed the view that they no longer need to provide learning materials themselves, as students will seek them out elsewhere. Instead, the teacher sees their role as guiding this process, acting as role models, evaluating, and making sense of the information they find, describing this shift based on their own classroom experience. In this regard, one teacher described: "*Even when they [the students] Google something, they still turn to the real, flesh-and-blood teacher and ask, 'Does this seem right?' or 'What do you think about it?' So they still need that confirmation and validation from the teacher to feel assured that what they're doing is actually correct.*" (T20). Another teacher agreed and elaborated: "*I get the sense that, at the same time, exchanging opinions and asking, 'What do you think about this?' or 'Where do you stand on that?' is actually becoming more important, which I find quite interesting. In my*

experience, students are increasingly seeking orientation or guidance. As we already mentioned with Googling, the internet is vast, and AI is an even larger and more complex space if you think of it as a kind of landscape. I believe young people need direction, and I've definitely noticed that, perhaps it's also a result of the current crisis-ridden times, but it all seems interconnected. This is also shifting the teacher's role." (T5).

While affirming the teacher's responsibility to serve as a role model for students, one teacher expressed skepticism toward the term learning facilitator by stating: "*There's often an underlying question: shouldn't teachers now become learning facilitators rather than traditional instructors? I find that discussion can sometimes feel very technical. After all, we're always present as individuals, and each teacher relates to students differently, some more closely, some with more distance. Our personalities themselves have an impact on how we engage with students. That's why I think it's important to be cautious and not fall into overly ideological or technocratic thinking. Saying, "We're no longer teachers; we're just facilitators now," to me, that misses the point. I've always accompanied my students on their learning journey.*" (T5) This perspective underscores that, while roles may evolve, the essence of teaching lies in the personal engagement and guidance teachers provide as role models, which cannot be reduced to a purely technical or facilitative function.

Teachers as Enlighteners

Other teachers reported a growing sense of responsibility to educate their students on topics such as social media, social values, and fake news, competencies that go beyond factual knowledge. In doing so, they assume a role that can be characterized as that of an enlightener, by highlighting AI's threats, such as filter bubbles and fake news online, and promoting social values.

Teachers who view their responsibility as **highlighting AI's threats, such as filter bubbles and deep fakes**, describe: "*it's more about taking on responsibility in the sense of feeling accountable for showing students what AI can and […] there's an added layer of responsibility involved. […] For example, with deepfakes and all the related developments, it's essential to teach students how to verify sources and use reliable information. And yes, that need has certainly increased with the rise of AI. Especially when it comes to deepfakes, where voices can be imitated easily, and entire avatars can be created that look exactly like you, students really need to be made aware. If they see a video where it looks like a friend is doing something inappropriate or problematic, and it looks and sounds exactly like them, they need to learn to pause and think: Could this really be true? Is this real? It's not like in the past, where people would assume a video must be real because it was too hard to fake. Now we also need to question things like: where does ChatGPT get its sources from? Is the information it provides actually accurate and reliable? That's definitely an*

added responsibility we now have." (T2) Other teachers highlighted challenges such as uneven levels of media literacy among students. As one explained: "*We have modules on topics like data protection, social media, and fake news. But I believe this is an area we need to keep working on continuously. The level of awareness among students varies greatly. Some are well-informed and highly reflective, but there are still many who simply say, 'I read that somewhere, so it must be true.' That's why educating students about how filter bubbles form and how certain algorithms reinforce their existing opinions is so important. I think our school is on a good path in this regard, depending on the subject area. However, I believe this will continue to be an integral part of teaching and is not something that will diminish over time*." (T13) Concerns about responsible GAI use were also raised. One teacher observed: "*I definitely feel that the sense of responsibility has increased. I often notice that students handle AI quite carelessly. For example, I've seen screenshots where the AI says something like, 'Oh, this is the assignment for Ms. M., she likes it when you use these particular vocabulary words' That's when I start to become concerned, issues like data privacy and similar matters come up. Especially in that area, I do feel that we, as teachers, have more responsibility now than we did before*." (T 15). This was echoed by another, who framed the issue succinctly: "*We have the responsibility, so to speak, to distinguish between truth and falsehood, and, in any case, to educate students about how these technologies work*." (T25)

Other teachers reflected on the need to **promote social values** by sharing a story, suggesting that new professional role responsibilities may also arise from broader societal shifts in values: "*Sixty years ago, as I've already told my students, people who stubbornly insisted on their opinions without a solid foundation were ridiculed by society. Today, it's the opposite, when someone presents facts, they're the ones who get laughed at. This shift, in this particular form, is something we need to address when working with AI, to ask what dangers it ultimately poses for society, for the individual, and for us as a class.*" (T28) These accounts highlight the growing importance of promoting social values in the age of GAI, which reshape teachers' professional role responsibilities toward fostering ethical awareness and resilience against misinformation. Likewise, a teacher reflected on a notable shift in schools' approach to political topics. In the past, such issues were often avoided; however, there is now a growing recognition of the need to actively engage with students to uphold and promote core societal values in an era of widespread misinformation: "*In the area of democracy and values education, it's important to consciously create regular time slots that allow current political topics to be brought into the classroom. In recent years, we actually tried to keep those topics out. Today, we realize that 'no, it's actually good to take a stand and bring these issues into the discussion.' It's part of returning to a more holistic pedagogical role that teachers are meant to fulfill.*" (T11)

Teachers as Challengers

Another group of teachers sees their primary role responsibilities in continuously challenging students to improve and is therefore categorized as "challengers." They are characterized by viewing their role responsibilities in promoting foundational GAI literacy, identifying areas where students can improve and grow, and critically engaging with GAI-generated content collaboratively with the students.

Those who described their role responsibilities as challenging students in the context of increased GAI use expressed a generally positive attitude toward GAI augmenting individuals capabilities, aligning with the view that "*Students need to use it, but they also have to understand what comes out of it and be able to follow the process that led there*" (T4). This emphasizes the importance of **promoting foundational GAI literacy**, as one teacher argued: "*I think things will increasingly move in the direction of AI, and that certain professions will essentially be replaced by AI. As a result, students in the 21st century will need to be able to program or use AI accordingly. And in that sense, I'm preparing them for their professional future*." (T1, Pos. 2) One teacher described their approach in doing so: "*And we have to show them, okay, it's great that you used AI here, we're not fundamentally against that, but if you want to get a good result out of it, you still need to do this and that*" (T24). GAI literacy, then, is not just about technical competence. It is about cultivating students' capacity to question what they read, see, and produce. "*They should also learn to evaluate and critically question their self-generated answers, just as they should evaluate and question what is presented to them elsewhere, whether on social media, in the press, or even in film and television, where AI is now also being used*" (T19). This shift in required competencies was also reported to redefine assessment. "*Assessments are sometimes not just about subject knowledge, they also involve a pedagogical aspect. And as a teacher, you definitely need to take another look and critically review the whole thing*" (T6)

This shows, challengers also see their responsibilities in **identifying areas where students can improve and grow** by stating that "*assessments actually take more of a back seat, and the focus is really on helping the student improve. It's more about giving suggestions for improvement and supporting a solid revision of their own work*." (T3) This issue reflects a broader educational challenge, namely the increasing need to shift focus from content delivery to the development of transversal skills. As one teacher observed: *" I mean, I can put together a presentation really quickly now, that's three clicks and a two-minute wait. And if I have a bit of money, it'll look even nicer if I get a paid account, depending on my background. But does that mean I can actually present it? Does that mean I know what I'm talking about? These are all aspects that are now coming much more into focus."* (T5) Accordingly, teachers are becoming more focused on fostering soft skills. As another teacher puts it, "*I believe I now have more responsibility for fostering soft*

skills, things like critical thinking, encouraging students to engage deeply with topics instead of just scratching the surface. Organizational skills in students, too. I think these aspects will play an even greater role in the future." (T1).

Another teacher emphasizes their responsibility in **promoting critical engagement with GAI-generated content collaboratively**. The teacher encourages students to use GAI, but not blindly: *"I do recommend that they [the students] use it [GAI]*" (T15), he says, *"but not just in the sense of: 'Write the analysis for me, in perfect English and without errors,' because then I don't learn anything from it. I might have the homework done for the next class, but in the exam I still have to be able to do something on my own."* (T15) Instead, the teacher urges "*to reflect on its use. What do I actually need it for? How is it supposed to help me? And how can I use it in a way that still results in learning gains or has an effect that helps me progress and develop further on my own?"* (T15), guiding students toward a deeper engagement with the content that promotes learning gains and personal development. Yet, the development in classrooms is mixed. Another teacher notes: *Some students have actually become quite good at prompting, I think. Probably better than me in many cases. Others are still very rudimentary in their approach and then wonder why the result doesn't match what was actually required. But often, they don't even notice. They just copy and paste, what matters is that something is written there. Unfortunately, that kind of critical reflection is often missing."* (T9, Pos. 2)

5 Discussion

By exploring the changes in professional role responsibilities among secondary teachers through the increased use of GAI in education, this study shows that both the use and perception of GAI vary considerably, and with them, the responsibilities teachers associate with their professional roles (Harder Fischer et al., 2024). In the context of democratized AI integration (Krakowski, 2025), where adoption can emerge both top-down from institutions and bottom-up from teachers or students, these adaptations highlight the continuous changes and ongoing readjustments. The analysis identified four distinct professional role responsibility types, each representing a particular configuration of GAI perception and associated job crafting practices that provide insights on how AI is altering the teacher-student interaction (Gentile et al., 2023). While the scope of this study does not extend to explaining full transformations in teachers' professional role identities, the observed shifts in responsibilities can be understood as partial expressions of identity work, in which teachers negotiate the meaning of their role in light of GAI's affordances. From these patterns, four propositions emerge, offering a foundation for future research to examine how differing perceptions and uses of GAI shape the evolving nature of teachers' professional role responsibilities. In doing so, this study provides empirical insights into how teachers envision diverse digital

futures, thereby responding to the call by Schlagwein et al. (2025) for research that acknowledges plural futures in uncertain environments

The following section presents this study's three key contributions to IS research.

First, the findings indicate that in the sense of job crafting, GAI is predominantly perceived and utilized as a tool for the reproduction of familiar tasks, suggesting that its full technological potential remains underexploited. For many teachers, their own workflows are largely conceived as separate from the GAI, following a linear process in which they provide a prompt, receive the GAI-generated output, and subsequently revise the material offline. This interaction pattern reflects a transactional (in the sense of automation) rather than an iterative engagement (linked to augmentation) with the technology. However, there is significant scope to increase awareness among teachers of the possibilities afforded by more dynamic GAI use. By engaging in iterative prompting, that is, refining prompts, providing additional input, and deliberately steering the GAI towards more sophisticated or targeted outputs, teachers could, and some already do, begin to co-create content with the system, effectively "training" it for their pedagogical purposes. Such practices move beyond mere substitution or automation towards augmentation, where the GAI becomes an active partner in enhancing the quality, adaptability, and creativity of teaching materials and processes.

Second, the perception of GAI and its associated use patterns appear to influence how teachers construe the responsibilities inherent to their professional role. This dynamic can be understood through the lens of job crafting, which describes how professionals actively shape and redefine their work to align with their own preferences, motives, and passions (Tims et al., 2012). The findings indicate that teachers who perceive GAI primarily as a supportive tool are more likely to maintain traditional conceptions of their role responsibilities, like being consultants or cheerleaders, by motivating their students. In contrast, teachers who adopt GAI as a sparing partner or critical friend, in the sense of augmentation, may reconceptualize aspects of their role responsibilities, for instance, challenging students and enlightening GAI's impact on outcomes and society. Consequently, this study suggests four propositions that future research should quantitatively examine:

- **Proposition 1**: *Teachers who view GAI primarily as a supportive tool tend to perceive their primary professional role responsibilities in motivating students*. In this sense, they perceive a risk that students may lose the motivation to learn independently. As a result, they focus on encouraging students to think critically, persist in their learning, and develop problem-solving skills, even when GAI provides quick or seemingly complete answers. This orientation reflects a commitment to

preserving students' intellectual autonomy and intrinsic motivation within an GAI-rich learning environment.

- **Proposition 2**: *Teachers who experience GAI as a disruptive presence are inclined to see their responsibility as enlightening students about GAI's ethical implications and societal impact.* In this context, GAI is often perceived as having significant societal impacts, with references to issues such as deepfakes. This perspective emphasizes preparing students to navigate GAI's influence responsibly by fostering ethical awareness and reflective judgment.
- **Proposition 3**: *Teachers who perceive GAI as a critical friend are inclined to see their responsibility in consulting students to effectively use GAI in their learning processes*. More precisely, teachers who view GAI as helpful for tasks such as translation when uncertainties arise or for providing differentiated content perceive increased responsibilities in guiding students through their learning processes and GAI use, while also serving as role models.
- **Proposition 4**: *Teachers who engage with GAI as a sparring partner tend to perceive their primary professional role responsibilities in challenging GAI's reliability and scrutinizing its outputs.* This may be particularly evident among teachers who engage in conversations with chatbots to challenge their ideas. As a result, they may perceive increased role responsibilities in critically evaluating GAI outputs and facilitating problem-based learning, thereby ensuring that autonomy remains with the GAI user.

These role responsibility adjustments align with role identity theory, which posits that professional self-concept is shaped by the enactment of responsibilities within a given role. While this study does not seek to explain full transformations in teachers' professional role identities, the shifts observed here can be seen as expressions of identity work in progress, partial but significant redefinitions of "what I do" and "what my responsibilities are" as a teacher in the era of GAI. By connecting these findings to job crafting and role identity theory, this study extends IS research by demonstrating that GAI adoption in education is not merely a matter of technology implementation, but a "sociomaterial entanglement" (Leonardi, 2012) in which teachers' practices, identities, and technologies continually shape one another. This perspective invites future research to examine the longitudinal interplay between GAI-related job crafting and evolving professional role responsibilities, and to identify the conditions that enable these changes to enhance rather than diminish teachers' sense of agency, meaning, and professional purpose.

Third, the findings also have implications for practice as they indicate that teachers perceive an increase in their professional role responsibilities as a result of GAI integration. These expanded responsibilities include ensuring that students maintain and develop their cognitive skills rather than becoming over-reliant on technology, providing education on issues such as data privacy and the risks of deepfakes, and fostering

critical reflection alongside an appreciation for the value of factual accuracy. While these roles position teachers as key mediators in responsible GAI use, they also introduce additional cognitive and emotional demands. Future research should therefore examine the nature and extent of this associated workload, with the aim of identifying strategies and support mechanisms that can help mitigate these burdens and enable teachers to fulfil their expanded role responsibilities effectively.

6 Conclusion

This study aims to shed light on how secondary teachers experience GAI as it becomes an integral part of their professional landscape. By revealing four distinct ways in which teachers craft their jobs and use GAI (supportive tool, critical friend, disruptive presence, and sparring partner) and linking these to corresponding professional role responsibility types (consultants, cheerleaders, enlighteners, and challengers), it illustrates that GAI's role in education is not fixed but continuously negotiated in practice. These findings highlight that the future of teaching will not simply be shaped by GAI's capabilities, but by the evolving interplay between teachers' perceptions of GAI, their corresponding professional role responsibilities, and the sociomaterial entanglements that GAI brings to the classroom.

As GAI continues to advance, the challenge for researchers, policymakers, and practitioners is to cultivate conditions in which these teacher-GAI collaborations strengthen, rather than diminish, the agency and professional purpose of educators. The four propositions outlined in this study provide a foundation for further inquiry into how educational work can and should be meaningfully reimagined in collaboration with GAI. Ultimately, understanding teachers' digital futures and shaping this transformation is essential, not only for the teachers of today but for the professionals of tomorrow whose learning will be shaped by their guidance.

References

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, *15*(3). https://eric.ed.gov/?id=EJ1395102

Aldeman, N. L. S., De Sá Urtiga Aita, K. M., Machado, V. P., Da Mata Sousa, L. C. D., Coelho, A. G. B., Da Silva, A. S., Da Silva Mendes, A. P., De Oliveira Neres, F. J., & Do Monte, S. J. H. (2021). Smartpathk: A platform for teaching glomerulopathies using machine learning. *BMC Medical Education*, *21*(1), 248. https://doi.org/10.1186/s12909-021-02680-1

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. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

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), 10.

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology*, *13*, 836650.

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success. *Information Systems Frontiers*, *25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Gentile, M., Città, G., Perna, S., & Allegra, M. (2023). Do we still need teachers? Navigating the paradigm shift of the teacher's role in the AI era. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1161777

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

Harder Fischer, L., Nicolajsen, H. W., Marttila, S.-M., & Sandbukt, S. (2024). *Crafting Meaningful Generative AI-Enabled Knowledge Work*. https://aisel.aisnet.org/ecis2024/track01_peoplefirst/track01_peoplefirst/10/?trk=public_post_comment-text

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

Hinds, P., & Von Krogh, G. (2024). Generative AI, Emerging Technology, and Organizing: Towards a theory of progressive encapsulation. *Organization Theory*, *5*(4), 26317877241293478. https://doi.org/10.1177/26317877241293478

Hovorka, D. S., & Mueller, B. (2025). Speculative foresight: A foray beyond digital transformation. *Information Systems Journal*, *35*(1), 140–162. https://doi.org/10.1111/isj.12530

Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(5), 87–101.

Jaiswal, A., & Arun, C. J. (2021). Potential of Artificial Intelligence for transformation of the education system in India. *International Journal of Education and Development Using Information and Communication Technology*, *17*(1), 142–158.

Johnson, D. G., & Verdicchio, M. (2017). AI Anxiety. *Journal of the Association for Information Science and Technology*, *68*(9), 2267–2270. https://doi.org/10.1002/asi.23867

Jones, M. (2014). A matter of life and death. *MIS Quarterly*, *38*(3), 895-A6.

Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization*, *35*(1), 100560.

Lameras, P., & Arnab, S. (2021). Power to the teachers: An exploratory review on artificial intelligence in education. *Information*, *13*(1), 14.

Leonardi, P. M. (2012). Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they different? Do we need them. *Materiality and Organizing: Social Interaction in a Technological World*, *25*(10), 10–1093.

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

Orlikowski, W. J. (1992). The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Science*, *3*(3), 398–427. https://doi.org/10.1287/orsc.3.3.398

Raskin, P., Banuri, T, & Gallopin. (2002). *Great Transition: The Promise and Lure of the Times Ahead*.

Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. *MIS Quarterly*, *37*(4), iii–xviii.

Schlagwein, D., Currie, W., Leimeister, J. M., & Willcocks, L. (2025). Digital futures: Definition (what), importance (why) and methods (how). *Journal of Information Technology*, *40*(1), 2–8. https://doi.org/10.1177/02683962241301544

Stets, J. E., Burke, P. J., Serpe, R. T., & Stryker, R. (2020). Getting Identity Theory (IT) Right. In S. R. Thye & E. J. Lawler (Eds.), *Advances in Group Processes* (Vol. 37, p. 0). Emerald Publishing Limited. https://doi.org/10.1108/S0882-614520200000037007

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*, *22*(2). https://doi.org/10.17705/1jais.00663

Tims, M., Bakker, A. B., & Derks, D. (2012). Development and validation of the job crafting scale. *Journal of Vocational Behavior*, *80*(1), 173–186. https://doi.org/10.1016/j.jvb.2011.05.009

Yusuf, A., Pervin, N., Román-González, M., & Noor, N. M. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education*, *12*(2), e3489. https://doi.org/10.1002/rev3.3489

Zhai, X. (2024). Transforming Teachers' Roles and Agencies in the Era of Generative AI: Perceptions, Acceptance, Knowledge, and Practices. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-024-10174-0

Appendix

Appendix I. Semi-structured interview guideline

Introduction

First of all, thank you very much for taking the time to participate in this interview. The aim of this interview is to explore your experiences and perspectives on the use of AI applications in your teaching and school-related practices. There are no right or wrong answers, what matters is your individual viewpoint regarding your use of AI tools such as ChatGPT, DeepL, or Perplexity.

The interview will take approximately 20 minutes. It will be audio-recorded and stored for the purpose of analysis. The interview will be transcribed and subsequently analyzed. The audio files will be deleted after transcription. All descriptions of the analysis results will be fully anonymized.

All data collected during the interview will be processed in accordance with applicable data protection regulations and used exclusively for the stated study. Your responses will be treated with strict confidentiality, and participation is voluntary.

ID block	Questions	Conceptual basis
1	• Could you briefly describe your professional background? • How long have you been working as a teacher? • What is your subject area or areas of expertise as a teacher?	Demographic data
2	• What do you consider to be AI applications?	Understanding of AI
3	• In which activities or areas do you already use AI applications? And why? Please mention 2–3 examples.	Fields of AI application
4	• Has your understanding of your role responsibilities as a teacher changed through the use of AI in your teaching? If so, how?	Impact of AI on responsibility and role identity
5	• To what extent should responsibility for AI-supported decisions be shared between teachers and AI systems? • How much responsibility can you take for AI-generated content? • Where do you draw the line between your responsibility and that of the AI system, for instance, in the case of incorrect decisions?	Boundaries of Responsibility

6	• What ethical concerns do you have regarding the use of AI? • To what extent do these concerns influence how you use AI in the classroom?	Self-reflection and ethical responsibility
7	• What measures do you consider necessary to ensure the responsible use of AI applications?	Support measures

Appendix II. Participants

Reference	Gender	Subject
T01	female	Economics
T02	male	Economics
T03	male	Politics and Society & German
T04	male	Economics
T05	female	mathemathics & german
T06	female	Business Administration & English
T07	female	Business Administration & Accounting
T08	male	Business Administration & English
T09	female	German, Biology, Religion
T10	male	Business Administration, Economics & Accounting
T11	male	Computer Science
T12	male	Medicine & Sports
T13	male	English
T14	male	Economics
T15	female	German & English
T16	male	Economics & English
T17	female	Economics & Religion
T18	female	German & English
T19	male	Business Administration
T20	male	Economics
T21	female	English & German
T22	female	Economics
T23	female	Economics
T24	female	Insurance & Business Administration
T25	female	Business Administration
T26	male	Biology
T27	female	Business Administration

T28	male	Religion

Appendix III. Overview of Data Structure – AI use in teaching tasks

1st-order constructs	**2nd order themes**	**Overarching dimensions**
"generating case study" "Create an exercise" "created quiz" "a bingo game"	Developing practice exercises	AI as a supportive tool
"generate feedback text modules"	Provide learner feedback	
"I often give it my own materials and have it rephrase them nicely" "change the format"	Producing informational materials	
"performance test" "multiple-choice questions"	Creating assessments and exams	
"how to structure a unit on topic XY"	Structuring lesson plans	
"visualize content" "[AI help me to produce] anything artistic"	Visualizing content through AI-generated images	AI as disruptive presence
"podcast instead of just reading a text"	Producing educational podcasts	
"instructional videos" "get a script for a video"	Producing dynamic learning videos	
"have a conversation partner" "someone to exchange ideas with" "How would you do it?" "I get input that's valuable to me" "brings in different perspectives from the entire data pool" "sharpen one's thinking" "always having a second perspective"	Co-creating ideas with AI as thinking partner	AI as sparring partner
"get ideas" "ask about certain things" "obtaining information even more easily	Searching and retrieving instructional content	AI as a critical friend

“because I often ask the AI when I have a certain topic”		
“the aspect of differentiation that AI greatly simplifies.” “design tasks that are sometimes easier, sometimes more difficult” “or rephrase it in a simpler way”	Delivering differentiated instructions	
“for working with English” “It then shows me that, yes, you can write it that way”	Translating learning materials	

Appendix IV. Overview of Data Structure – Professional Role responsibilities

1st-order constructs	**2nd order themes**	**Overarching dimensions**
“that feedback is simply motivational” “really need a lot of support and, let’s say, even a bit of a push”	Motivate	Cheerleader
“on building relationships with them” “the student-teacher relationship”	Foster interpersonal relationships	
“showing students what AI can” “students really need to be made aware” “issues like fake news and hate speech” “topics like data protection, social media, and fake news” “filter bubbles” “students handle AI quite carelessly” “issues like data privacy” “to speak, to distinguish between truth and falsehood”	Highlight AI’s threats	Enlightener
“teach students how to verify sources and use reliable information” “to ask what dangers it ultimately poses for society”	Foster civic responsibility	
“values education”	Promote social values	
“talk to students about learning how to learn” “It doesn’t help to just let them loose with it” “learning spaces” “We’re learning coaches”	Guide students	Consultant

“learning facilitator” “student-centered form of teaching” “student-centered form of teaching”		
“confirmation and validation from the teacher” “students are increasingly seeking orientation or guidance”	Serve as role model	
“helping the student improve” “It’s more about giving suggestions for improvement” “supporting a solid revision of their own work” “have more responsibility for fostering soft skills”	Identify areas for improvement	Challenger
“to reflect on its use” “critical reflection” “What do I actually need it for?” “How is it supposed to help me?” “things like critical thinking” “They should also learn to evaluate and critically question their self-generated answers” “involve a pedagogical aspect”	Encourage critical thinking	
“they also have to understand what comes out of it” “be able to follow the process that led there” “be able to program or use AI accordingly”	Promote AI literacy	

Paper VI

Passlack, N., Hammerschmidt, T., & Posegga, O. (2025). With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?. *Information Systems Frontiers (ISF)*

https://doi.org/10.1007/s10796-025-10648-5

Information Systems Frontiers
https://doi.org/10.1007/s10796-025-10648-5

With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?

Nina Passlack[1] · **Teresa Hammerschmidt**[1] · **Oliver Posegga**[1]

Received: 4 October 2024 / Accepted: 27 August 2025
© The Author(s) 2025

Abstract
The growing autonomy and self-learning abilities of advanced technologies alter the dynamics of human-AI interaction, while also raising important ethical concerns. This study calls for rethinking AI literacy to encompass additional abilities necessary for responsible human-AI interaction. Beyond enhancing efficiency and performance, research on AI literacy must address unintended consequences and risks, and thus requires a digital responsibility perspective. Through interviews with different groups of knowledge workers, the study outlines factors that enable and influence knowledge workers to interact responsibly with AI technologies. Our research uncovers tensions concerning the concept of AI literacy that may challenge responsible human-AI interaction. These concern especially problems related to skill development and responsibility attribution. These problems emphasize the need for shared responsibilities and to ensure alignment between knowledge workers' abilities and their roles – what we call the "role-ability fit." We further offer recommendations to promote responsible human-AI interaction.

Keywords AI literacy · Human-AI interaction · AI knowledge · Responsibility

1 Introduction

Digital technologies based on artificial intelligence (AI) are widely used in knowledge workers' daily lives (Dwivedi et al., 2023); their growing power reminds us of the old adage that "with great power comes great responsibility." This growing power stems from AI's increased cognitive and self-learning abilities to generate knowledge (Berente et al., 2021), prepare decisions, and carry out tasks that were previously done by individuals (Rai et al., 2019), thus transforming aspects of contemporary society and business life (Constantinescu et al., 2021). The greater responsibility arises from the blurred lines of responsibilities when AI technologies operate (semi-)autonomously, and accountability for unintended consequences remains unclear (Giermindl et al., 2022; Matthias, 2004). A lack of the skills required for interacting with AI may lead to gaps between individuals who are and are not able to use AI technologies responsibly (Celik, 2023). It may also result in unintended consequences that create inequalities, such as the exclusion of marginalized groups of people (Alexopoulou et al., 2022; Celik, 2023).

To avoid such inequalities and mitigate potential negative or unintended consequences of human-AI interaction, we need to rethink the concept of AI literacy. First, individuals must acquire new abilities that will enable them to interact with AI responsibly (Rana et al., 2022). These include being able "to evaluate AI technologies critically; communicate and collaborate effectively with AI; and use AI as a tool" (Long & Magerko, 2020, p. 2). Current research refers to this as AI literacy (e.g., Ng et al., 2022; Pinski & Benlian, 2023), and some valuable conceptualizations of AI literacy exist (e.g., Cetindamar et al., 2022; Heyder & Posegga, 2021; Pinski & Benlian, 2023; Wang et al., 2022). Second, conceptualizations of AI literacy should take into account the various influencing factors that shape skill development and determine competency requirements. Building upon existing conceptualizations, our work aims to outline components that are critical to AI literacy for achieving *responsible* outcomes of human-AI interaction alongside those that are essential for achieving *efficient* outcomes, such as increased performance and task efficiency.

✉ Nina Passlack
nina.passlack@uni-bamberg.de

[1] Information Systems and Social Networks, University of Bamberg, An Der Weberei 5, 96047 Bamberg, Germany

Published online: 22 October 2025

While efficient outcomes focus on maximizing utility related to time, finance, and other resources (Reinhardt, 1992), responsible human-AI interaction requires consideration of the potential consequences of human-AI interaction that can produce unintended side effects (Kegel & Ghanem, 2024) with respect to employee well-being, fairness, other social implications, and so on. While these responsibility dimensions may not directly affect efficiency in the short term, they may in the long term (Benlemlih & Bitar, 2018). Recognizing the need for organizations to remain competitive and financially sustainable, we acknowledge the crucial role of efficiency in enabling responsible human-AI interaction. Yet, we argue that focusing mainly on AI literacy that facilitates process efficiency in human-AI interaction is a one-sided approach. We suggest instead that knowledge workers must balance multiple factors to ensure responsible outcomes in human-AI interaction. Since inequalities may exist that divide certain groups of people from others, as outlined by Zamani and Vannini, we argue that a lens of digital responsibility is essential when conceptualizing AI literacy, in order to prevent such unintended consequences.

This requires a digital responsibility perspective to mitigate the lack of abilities required for responsible human-AI interaction. Consider, for instance, the introduction of AI in organizations with a focus on training knowledge workers in skills to make quicker and more precise decisions, which could make work practices more effective in the short term (Mayer et al., 2020). Not preparing those same knowledge workers profoundly for AI's impact on their work practices might, for instance, lead them to see AI as a threat to their identity in the long term (Mirbabaie et al., 2022; Parker & Grote, 2022). People may have difficulties understanding the processes behind AI and have issues defining their roles in the context of human-AI interaction (Benbya et al., 2021). This issue becomes particularly concerning when individuals are unaware that using AI can lead to moral violations. For example, medical AI used for diagnostics in healthcare may be trained on datasets that are not representative of all populations, potentially resulting in higher misdiagnosis rates for certain ethnic groups (Lebovitz et al., 2022; Yokoi et al., 2021). Thus, responsibility gaps can occur when it comes to unintended outcomes of human-AI interaction (Smith & Vickers, 2021), requiring a more conscientious approach that pays greater attention to what fosters *responsible* human-AI interaction. We aim, therefore, to answer the following research question:

What Shapes AI Literacy for Responsible Interactions of Knowledge Workers with AI?

After synthesizing existing conceptualizations of AI literacy and outlining the different aspects, building blocks, related literacies, and analysis levels of AI literacy in our theoretical background, we conducted qualitative interviews ($n=28$) with five different groups of knowledge workers to account for different professional perspectives on human-AI interaction and thus answer the research question. Our key findings highlight tensions that challenge the conceptualization of AI literacy through the lens of digital responsibility and the examination of factors that influence responsible human-AI collaboration. Our analysis does not claim to provide an exhaustive account of all components relevant to AI literacy; rather, its primary aim is to offer a perspective on rethinking the concept of AI literacy by highlighting which constructs become relevant when the objective is not only effective but also responsible human-AI interaction. We emphasize moving beyond a narrow focus on efficiency, highlighting the need for AI literacy to equip knowledge workers not only with technical skills but also with the capabilities to engage with AI systems in an ethically sound, accountable, and context-sensitive manner and thereby consider possible influencing factors that shape responsible interaction with AI.

While different perspectives exist, our interviews emphasize the importance of users' personal dispositions, human-centered aspects, ethical literacy, and the factors influencing AI literacy development across micro-, meso-, and macro-levels, all of which are crucial to consider for fostering responsible human-AI interaction. Finally, we emphasize the critical importance of achieving what we call the "role-ability fit" – the alignment between the responsibilities assigned to individuals in specific roles and their capacity to act responsibly and be held accountable for those duties. This insight leads us to underscore the growing relevance of "multi-literacy" – the ability of knowledge workers to navigate and understand diverse domains, enabling them to engage responsibly with AI technologies. Furthermore, given that AI technologies involve multiple stakeholders across different phases of development and application, establishing shared responsibility is indispensable. We suggest, only through this collective responsibility can we ensure that human-AI interactions are ethically sound and socially beneficial.

We see our work contributing to the field of Information Systems (IS) in two key ways. *First*, we contribute to IS research by enhancing previous conceptualizations of AI literacy with an empirical analysis of components that become increasingly relevant to dealing with AI technologies *responsibly*. Thereby, we rethink the concept of AI literacy through a lens of digital responsibility and investigate knowledge workers' perceptions regarding the evolving importance of abilities in the context of AI-driven transformation, with particular focus on responsible human-AI interaction. Our research thereby outlines how perceptions

may vary across knowledge workers and how understanding the concept of AI literacy as a modular concept may allow us to account for these diverse roles and perspectives. This approach also helps to identify which components of AI literacy may become particularly critical in ensuring responsible human-AI interaction. *Second*, our study outlines potential tensions in responsibility and influencing factors that impact the development of AI literacy among the workforce, which can be a challenge for the strategic management of AI in organizations. We outline problems that may hinder responsible human-AI interactions within organizational settings and propose possible strategies to support organizations in fostering responsible human-AI interactions.

The next section provides the theoretical background for our research. We then describe the methods used in our qualitative study. The fourth section presents our results, followed by a discussion of the problems identified. The paper ends with a conclusion section.

2 Theoretical Background

This section identifies prior conceptualizations of AI literacy (Section 2.1) and establishes theoretical foundations for understanding the importance of responsibility in human-AI interactions (Section 2.2). The procedure for reviewing the existing literature is detailed in Appendix 1.

2.1 AI Literacy, its Components, Related Literacies, and Possible Analysis Levels

Figure 1 presents an overview of potential AI literacy components relevant to people interacting with AI. Through our interviews, we aim to identify which of these components are essential for the responsible design, deployment, and use of AI technologies, as well as the management of human-AI interactions, to provide a holistic view of components relevant to knowledge workers interacting with AI. As Fig. 1 demonstrates, our findings from the literature consist of different *aspects*, *building blocks*, *technology-related literacies*, and different *analysis levels* that may influence the concept of AI-literacy and its development. Depending on the use case and context of human-AI interaction, specific combinations of these elements may become critical. Thus, we understand AI literacy as a toolbox, with each use case requiring a unique combination of specific aspects and building blocks, and being encouraged by other related literacies.

Literacy Aspects The academic discourse on literacy in the context of technology has proposed several classifications to distinguish different aspects of literacy. We categorized the factors into technical aspects – those directly related to understanding the AI – and human-centered aspects, which are also relevant in human–human interaction settings. For instance, McLean (2013) suggests that there are three

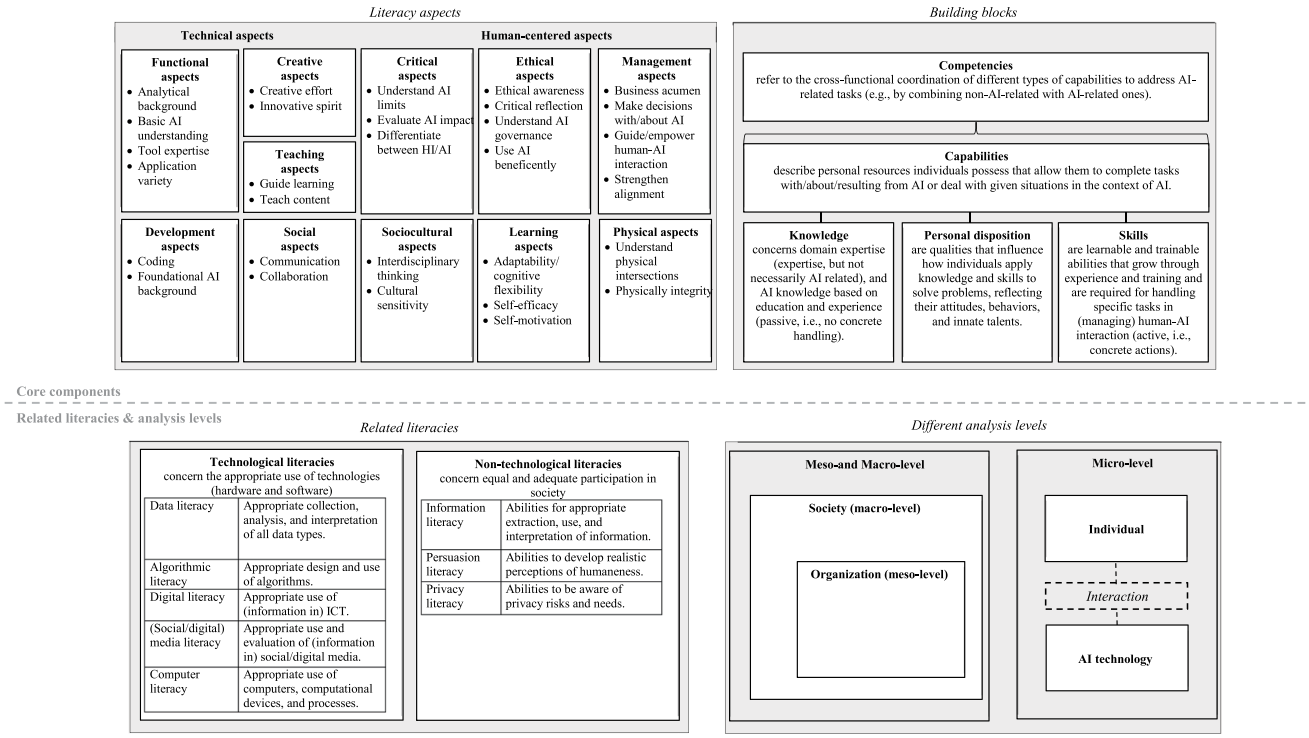


Fig. 1 Components of AI literacy

dominant literacy discourses that shape approaches to literacy in a technological context, which are differentiated into functional, critical, and socio-cultural. Adapting these notions, we propose that the *functional aspects* encompass the technical and analytical abilities required to manage data and use technologies effectively, such as basic AI understanding, which prior research has confirmed as key components of AI literacy (e.g., Carolus et al., 2023; Laupichler et al., 2022, 2023; Long & Magerko, 2020; Verma et al., 2022). The *critical aspects*, in turn, concern awareness of AI's impact on individuals, organizations, and society, as investigated, for instance, by Hermann (2022). Finally, the *socio-cultural aspects* focus on understanding the influence of cultural contexts on AI's outcomes, which is important to consider (Sanusi et al., 2022a, b).

Our review further highlights, however, that with AI advances and the partially uncertain consequences of human-AI interaction, additional aspects are becoming increasingly important and warrant further exploration. By *ethical aspects,* we refer, for instance, to the ability to recognize and navigate ethical conflicts and concerns, which has also been found to be part of AI literacy (Ng et al., 2021); yet, there is empirical research missing on how to foster those ethical aspects. By *learning aspects,* we refer to fostering one's interest in learning, being cognitively flexible, and having the ability to adapt dynamically (Martini & Sénéchal, 2012). While further research is required on how to "promote responsible … evolution of AI tools" (Dwivedi et al., 2023, p. 6), our focus is on understanding how learning aspects can provide support in this regard. S*ocial aspects* encompass communication and collaboration through and with AI, being identified for effective use of AI (e.g., Anton et al., 2020; Long et al., 2022), becoming particularly critical in case of unintended and unpleasant outcomes. As we suggest, *creative aspects* involve the identification of solutions to problems in the sense of problem solving (Ali et al., 2021; Anton et al., 2020; Su & Yang, 2023); more research is needed to explore how creativity can be applied when addressing responsibility issues. Finally, *physical aspects* are essential, as it is crucial to "understand that some AI systems can physically act on the world" (Long & Magerko, 2020, p. 6), such as in "robot-assisted surgeries" (Jaiswal et al., 2022, p. 1190). Given the broad spectrum of AI literacy components relevant to human-AI interaction, our empirical investigation aims to identify which aspects are most critical to ensuring responsible outcomes and to explore the factors that influence their development.

The literature also highlights the possibility of further literacy aspects that may be particularly relevant for specific technologies and groups of knowledge workers. For instance, we identified what we term *managerial aspects*, which encompass the ability of leaders and managers to recognize business opportunities and challenges arising from AI transformation, to make informed decisions grounded in business acumen, and to build trust while guiding their teams, all of which are essential for working with AI effectively (Alsheiabni et al., 2019; Anton et al., 2020). However, a gap remains in understanding how managerial dimensions are influenced by specific factors that hinder responsible human-AI interaction. *Developmental aspects* concern the ability to program and develop AI technologies, including preliminary ideas regarding how to do so responsibly, such as being attuned to biases in the data (Ng et al., 2021).

To summarize, AI literacy aspects are one component of AI literacy that requires deeper exploration to determine how these aspects can contribute to responsible human-AI interaction. Thus, our investigation seeks to explore how certain knowledge workers may possess specific literacy aspects that could foster responsible human-AI interaction.

Building Blocks Through the review, we found that AI literacy includes different building blocks, namely skills, knowledge, personal dispositions, capabilities, and competencies. Although these building blocks are often used synonymously, they represent distinct concepts that collectively form literacy (Bawden, 2008).

Skills are learnable and often "task-specific" as they grow with experience (Yamaguchi, 2018), and hence can be developed through education and experience (Gabric & McFadden, 2001). *Knowledge*, by contrast, refers to the awareness and understanding of facts and information being analyzed and is based on an individual's education and experience (Dalkir, 2013, p. 8). This means that "knowledge is information possessed in the mind of individuals" (Alavi & Leidner, 2001, p. 109). Knowledge can be what is often called domain knowledge, which refers to profound knowledge and expertise in a specific domain or context (Bradley et al., 2006), or meta-knowledge, such as knowing "who knows what" in the sense of "knowledge about knowledge" (Evans & Foster, 2011, p. 721). Some researchers (e.g., Hemment et al., 2023; Sanusi et al., 2022a, b; Su & Zhong, 2022) further underscore in the literature the importance of what we call *personal dispositions*, which "are affective or dispositional qualities, encompassing attitudinal, behavioral, and socio-emotional qualities of how disposed people are to apply knowledge and skills to solve problems or address issues of personal, social, or workplace-related interest" (Frezza et al., 2018, p. 155). Thus, personal dispositions also define "how much the learner makes the content and skills his or her own" (Frezza et al., 2018, p. 154). They may include interest or aptitude (Ferguson, 1960), but also implicit motivation, which constitutes a person's natural endowments

– innate capabilities that exist independently of the external environment (Ferguson et al., 2008).

Capabilities are linked to the deployment of knowledge and skills, as well as personal dispositions, to complete tasks or deal with given situations (Wahl & Prause, 2013); they arise from external forces and grow with experience and/or education. It should be noted that capabilities also include talent and aptitude as "a joint function of ability and interest" (Ferguson, 1960, p. 131). *Competencies*, by contrast, are the "cross-functional integration and coordination of capabilities" to address tasks (Wahl & Prause, 2013, p. 67); they build on capabilities by combining knowledge, skills, and attitudes.

The diverse array of building blocks that comprise AI literacy – ranging from skills, knowledge, and capabilities to competencies and personal dispositions – underscores the complexity of achieving responsible human-AI interaction. While existing theoretical frameworks provide a foundational understanding of these components, there remains a critical need for empirical research to determine which of these building blocks become particularly relevant in order to guarantee responsible human-AI interaction. Understanding which building blocks are most pertinent allows us to develop more targeted strategies for enhancing AI literacy for responsible interaction.

Related Literacies Our literature review reveals several interdependencies between AI literacy aspects and aspects of other forms of literacy, showing the importance of recognizing the range of literacies that may be related and therefore also relevant to AI literacy.

Prior literature further highlights connections between AI literacy and other literacies in a technology context. For instance, *data literacy* – the ability to collect, analyze, structure, and interpret data (Qin & D'ignazio, 2010) – can be considered a prerequisite for AI literacy (Anton et al., 2020; Jaiswal et al., 2022). *Algorithmic literacy* – understanding and appropriately utilizing algorithms (Ridley & Pawlick-Potts, 2021) – is further related to AI literacy in that it intersects with data literacy and digital literacy (Anton et al., 2020; Jaiswal et al., 2022). *Digital literacy* is the ability to interact effectively with digital tools such as information and communication technologies (ICTs) (Gilster, 1997; Reddy & Sharma, 2020). It, thus, includes *information technology (IT) literacy*, which also encompasses interacting with information embedded in various technologies (Reddy & Sharma, 2020). While not all researchers explicitly use the term "digital literacy," many suggest that it plays a significant role in AI literacy: some acknowledge overlapping areas (e.g., Anton et al., 2020; Wang et al., 2022), while others argue that digital literacy is a prerequisite for AI literacy (e.g., Cetindamar et al., 2022; Jaiswal et al., 2022; Long & Magerko, 2020; Markauskaite et al., 2022).

Information literacy and *media literacy* (including social and digital media literacy) are also connected to AI literacy (Hermann, 2022): (social or digital) *media literacy* is the ability to engage critically with information presented through various online media channels (Manca et al., 2021), while *information literacy* is the ability to identify, understand, and work with necessary information (Boichak et al., 2019; Rosenfeld, 2016). *Computer literacy* is the ability to demonstrate basic computer operation and application use skills (DiSessa, 2018, p. 4), and thus is also related to AI literacy. Most recent literature also introduces *privacy literacy* "as the ability to manage, protect privacy, and identify privacy risks" (Carolus et al., 2023, p. 5), while one paper introduces *persuasion literacy*, that can be defined as "understand[ing] how human-like features" can impact humans' perceptions of AI (Carolus et al., 2023, p. 5) as related literacy.

This shows that literacies are inherently interconnected, as being literate in any given situation often requires the integration of multiple aspects of different literacies (Lankshear & Lawler, 1987). Multi-literacy refers to the simultaneous application of various literacies within a single context (Luke, 1997). It can be understood as "emphasizing different kinds of …literacies to help [individuals] to become skilled at moving among them in strategic ways" (Selber & Selber, 2004, p. 24). In the context of responsible human-AI interaction, it is crucial to recognize the range of literacies that may also be relevant. Thus, AI literacy may be viewed as a multi-literate concept, with different use cases determining the particular literacies required for responsible human-AI interaction. Our empirical investigation explores distinct intersections of literacies to make individuals literate for responsible human-AI interaction.

Different Analysis Levels Human-AI interactions in organizations can be understood as socio-technical systems that outline interdependencies between humans, technologies, and organizational elements (Leonardi, 2012). As a consequence, human-AI interactions shape patterns of action in organizations (Cecez-Kecmanovic et al., 2014), and, as a consequence, AI literacy can also be influenced by factors at various levels of analysis. Analysis at a *micro-level* can provide insights into the aspects that influence individuals to interact effectively with AI technologies, recognizing that AI is transforming how individuals communicate and work (Long & Magerko, 2020). Thus, at the micro-level, factors from both the individuals and the technologies may influence requirements for literacy for responsible human-AI interaction. Analysis at a *meso-level* investigates AI literacy that enables organizations to deploy AI (such as Mikalef

and Gupta (2021) investigated). It comprises "the abilit[ies] of a firm to select, orchestrate, and leverage its AI-specific resources" (Mikalef & Gupta, 2021, p. 2). At a *macro-level,* AI literacy can be analyzed from a global perspective (Kathala & Palakurthi, 2024). The literature discusses, for instance, which skills are needed to recognize AI-generated deepfakes and misinformation that have an impact on society (Ali et al., 2021). However, further insights are needed into how factors across different levels of analysis influence knowledge workers' AI literacy in the context of fostering responsible human-AI interaction.

To conclude, Fig. 1 illustrates that AI literacy is a multi-faceted concept that can be analyzed on different levels, is composed of multiple interrelated components, and is closely related to other forms of literacies. While in some situations, specific literacy aspects become relevant to human-AI interactions, which is linked to the term upskilling (Morandini et al., 2023), in certain situations, other abilities may become less relevant or unlearned through increased reliance on AI, which can be linked to deskilling (e.g., Rafner et al., 2022). We aim to shed light on both upskilling and deskilling by discussing how this may foster or pose challenges to responsible human-AI interaction. Moreover, our interviews explore how to rethink the concept of AI literacy with respect to relevant components of AI literacy. In doing so, we understand AI literacy as a flexible toolbox, with different tasks and professional roles demanding a distinct mix of aspects and building blocks. It is further shaped by related literacies and influencing factors across different levels of analysis.

2.2 Heading from Efficient Toward Responsible Human-AI Interaction

Recent literature underscores the need for a shift from prioritizing efficiency to emphasizing responsibility in human-AI interaction, challenging traditional concepts of AI literacy. For instance, Oeldorf-Hirsch and Neubaum (2023) argue that while some people may possess high self-efficacy and cognitive skills for engaging effectively with social media algorithms, their literacy often falls short when considering the broader consequences. The authors highlight that specific groups of people, such as the LGBTQ+communities, have to "face extra challenges in terms of their literacy to make algorithms in constantly changing social systems work for them" (Oeldorf-Hirsch & Neubaum, 2023, p. 28). This observation suggests that AI literacy must extend beyond the efficiency of using algorithms to include an awareness of the social and ethical implications of their use to avoid discrimination and exclusion of certain groups of people (Celik, 2023). Oeldorf-Hirsch and Neubaum (2023) advocate for enhancing literacy by making users aware of potential biases, as well as for improving developer literacy to ensure the responsible design of social media algorithms and thereby avoid exacerbation of social inequalities. This shift reflects a broader move from focusing solely on efficiency toward integrating responsibility within the concept of literacy to balance efficiency and responsibility.

Similarly, Bankins et al. (2024) argue that current AI research is divided between studies that focus on the skills and knowledge necessary to maximize technological efficiency and those that examine the impact of AI on human workers, particularly regarding their self-confidence and emotional well-being. The prevailing emphasis on performance, characterized by increased efficiency and optimized resource use, often overlooks the responsible deployment of AI technologies (Passlack et al., 2024). For human-AI interactions to be truly responsible, it is essential to consider not only the immediate technical outcomes but also the broader, often unintended, consequences on human labor and societal well-being (Bankins et al., 2024).

In light of these discussions, the concept of **digital responsibility** gains increasing significance when considering the components that constitute AI literacy for responsible human-AI interaction. Digital responsibility encompasses the responsible design, implementation, and adoption of digital technologies, particularly during digital transformation efforts. More concretely, the principle of digital responsibility can be defined "as a fundamental and value-based normative requirement that motivates actors to attain responsible digital transformation" (Trier et al., 2023, p. 464). There is a growing need for empirical insights across various use cases of knowledge workers' engagement with AI to understand what shapes responsible interaction with AI and also how to responsibly address the potential harms of AI technologies while effectively harnessing their benefits (Pappas et al., 2023).

Several digital responsibility-related frameworks exist in the academic discourse. For instance, discourse around responsible AI emphasizes the design and implementation of AI technologies in real-world settings, guided by ethical principles such as fairness, explainability, and accountability (e.g., Arrieta et al., 2020; Trocin et al., 2023). The concept of corporate digital responsibility (CDR) advocates for human-centric approaches to manage human-AI interaction responsibly within organizations, emphasizing common values and standards that guide how digital technologies and data are developed and used (e.g., Kunz & Wirtz, 2024; Pappas et al., 2023; Trier et al., 2023). Such frameworks, though, are often related to single perspectives of responsibility and are often very abstract. Therefore, our research aims to explore on a meta-level how digital responsibility is and can be embedded within AI literacy investigations.

Through empirical interviews with different knowledge workers, we provide insights into the specific abilities required to interact with AI technologies responsibly and the arising tensions.

In this work, we further conceptualize digital responsibility in human-AI interaction as encompassing two key dimensions: the state of *being responsible* and *acting responsibly*. We suggest that acting responsibly involves the proactive consideration of potential unintended consequences during the interaction with AI, grounded in people's ability to reflect on their behavior and morally evaluate specific outcomes (Eshleman, 2014). Someone *acting responsibly* is expected to adhere to socially accepted and approved behavioral standards, aligning with the expectations placed on "accountable" actors as outlined by Lerner and Tetlock (1999). In contrast, *being responsible* pertains to the attribution of responsibility assigned per duty, and is often linked to a person's role (Hamilton & Hagiwara, 1992). "To be responsible for an outcome and to act responsibly may still be linked in the sense that a responsible person is prepared to accept the consequences of his or her actions, and thus acts in accordance with anticipated judgments of responsibility, rather than just be blamed or praised after the fact." (Nordbye & Teigen, 2014, p. 102). People who act irresponsibly could still be responsible. Indeed, it can be argued that they should feel and be held responsible, particularly for irresponsible behavior (Nordbye & Teigen, 2014).

Acting responsibly when interacting with AI, thus, necessitates certain abilities linked to the aspects introduced in Section 2.1; without these, people lack control over their actions and, consequently, it can be perceived as inappropriate to hold them accountable (Horneber & Laumer, 2023; Martin, 2019). For instance, AI systems used in recruitment processes, such as online assessments, may inadvertently discriminate against individuals with specific conditions, such as dyslexia. In such cases, human intervention is essential to ensure the fairness of the process (Malik et al., 2021). While AI can increase efficiency by reducing the need for human involvement in the selection process, exclusive reliance on AI for decision-making requires a clearer attribution of responsibility within the human-AI interaction. To ensure responsible practices within this context, we suggest that recruiters must develop AI literacy that enables them to recognize and address potential conflicts.

The state of **being responsible** is tied to the specific duties associated with individual roles or the accountabilities of knowledge workers as defined at the organizational level, which involves taking responsibility for achieving certain outcomes. As an example, managers can be made responsible for the effective implementation of AI within organizations (Alsheiabni et al., 2019). Individuals in such roles, however, may lack the necessary abilities associated with AI maturity, which constitutes a prerequisite for fully assuming the responsibility inherent in their positions (Alsheiabni et al., 2019). Thus, enhancing AI literacy is crucial as it can facilitate the behavioral changes necessary to fulfill these roles. Otherwise, responsibility gaps may emerge, particularly in relation to the unintended outcomes of human-AI interaction (Matthias, 2004; Smith & Vickers, 2021). Thus, we argue that responsible human-AI interaction requires an understanding of the complex nature of responsibilities that shape what it means for individuals to "be responsible" in these interactions. Given that various knowledge workers are involved in human-AI interaction, we discuss this complexity in our work.

Consequently, we suggest both dimensions – acting responsibly and being responsible – to be critical. For instance, while people may be held legally or morally responsible for certain AI operations, such as when pilots operate autonomous cockpits, the erosion of their decision-making authority and manual skills can create a disconnect between their responsibilities and their control over systems (Holford, 2022). Aligning knowledge workers' role-based responsibilities with their ability to act responsibly requires the *role-ability fit*, which we define as the alignment between individuals' prescribed responsibilities based on their roles (being responsible) and their actual AI literacy abilities required to fulfill those responsibilities (acting responsibly). There is a need to explore how the abilities of individuals and AI can contribute to fostering responsible human-AI interaction (Trocin et al., 2023).

To summarize, using the concept of digital responsibility in our work helps us by providing a lens to better understand the abilities that drive knowledge workers to interact with AI technologies responsibly and in a manner that proactively considers potential harms and maximizes benefits. It involves acting responsibly – by ensuring that knowledge workers have the ability to reflect on the unintended consequences of AI technology use – and being responsible, which entails taking responsibility based on one's role and duties. We investigate this based on our empirical insights from interviews with different groups of knowledge workers.

3 Research Design: Qualitative Interviews

We conducted semi-structured interviews with different groups of knowledge workers to deepen our understanding of what shapes *responsible* human-AI interaction that considers but goes beyond efficiency. Thereby, we aimed to extend the "evolving set of ideas that emerge …within and among scholarly communities" through "learning with and from practitioners" (Burton-Jones et al., 2021, p. 306). We deemed interviews to be a suitable method as they can

provide insights into individuals' thoughts, motivations, and experiences that cannot be obtained through quantitative data (Sarker et al., 2013). In doing so, we aimed to make sense of AI literacy for responsible human-AI interaction as an object. While prior research offers conceptualizations of AI literacy, our study aims to explore how people understand and perceive the promotion of responsible human-AI interaction. Interviews have proven to be a good approach for achieving a better understanding of social phenomena (Myers & Newman, 2007). Given that "responsible human-AI interaction" is an intangible construct that is difficult to quantify and measure, the interviews enabled us to examine the concept of AI literacy, in which predefined variables remain unclear and challenging to articulate. Furthermore, employing interviews acknowledges the complexity of combining the concept of responsibility with AI literacy, which is shaped by individual experiences rather than being reducible to numerical data. By conducting interviews, we were able to generate propositions that can now be tested using quantitative methods. The interviews helped us reflect upon different use cases by interviewing diverse professional groups (see 3.1), searching for AI literacy components required for interacting with AI in a responsible way, and investigating influencing factors such as potential challenges and tensions in establishing them.

3.1 Sampling

To ensure a cross-section of knowledge workers in different use cases of human-AI interaction, we deliberately selected participants from five stakeholder groups chosen to capture a broad range of perspectives on AI literacy, ensuring that insights reflect varied contexts and AI literacy abilities required across different professions. The stakeholder groups and their selection criteria are: HR professionals (H) responsible for hiring and training employees to develop AI-related abilities among the workforce, thus providing an operational perspective; managers (M) responsible for AI deployment in organizations, thus providing a strategic perspective; AI designers and data scientists (D) who are experts in designing, developing, training, and fine-tuning the underlying AI models and their applications, thus providing a developmental perspective; researchers and educators (R) who teach and train on relevant AI literacy abilities, thus providing an educational perspective; and sales and product managers of AI technology providers to businesses (S), who add a consumer-producer perspective. All our interviewees (see Appendix 2 for details) have a minimum of three years' experience in their fields and direct experience in AI-related tasks. Table 1 is an overview of the different stakeholder groups.

We recruited participants through targeted outreach via LinkedIn and email, selecting individuals based on their public professional profiles and their expertise in AI-related activities. Potential participants were provided with information about the research project, including its aims, expected contributions, and the interview structure. Data collection took place between March 2023 and September 2024. All interviews were conducted in German by telephone ($n=2$) or video call using Zoom or Teams ($n=26$), and included one interviewer and one person taking notes. The interviews, which lasted an average of 42 min, were recorded and transcribed, and subsequently translated into English. In total, we conducted 28 interviews.

3.2 Semi-Structured Interviews

We employed a semi-structured interview guide (Myers & Newman, 2007) to ensure that "standard procedures are used from one interview to another" (Crewswell, 2014, p. 194). The interview guide (see Appendix 3) was pre-tested on three pilot interviewees to ensure the clarity of the questions. The semi-structured interview guide has four main sections. In the first section, interviewees are asked about their background and working context, such as the work culture and the role AI plays for them at work. In the second part, interviewees are asked to describe different scenarios in which AI-based technologies are developed, sold, deployed, managed, educated about, or used in their business contexts. The third section focuses on what constitutes

Table 1 Interviewed groups of knowledge workers

Reference	Selection criteria	Perspective	N
H	HR professionals responsible for hiring and/or training employees (to become) familiar with the AI abilities the organization requires and who advocate for workforce AI literacy at the operational level	Operational perspective	5
M	Managers with decision-making responsibility for the implementation and use of AI in the organization and who may, therefore, provide a strategic perspective on AI literacy	Strategic perspective	6
D	Developers, data scientists, and AI specialists who use their AI knowledge to design/develop and/or use AI-based technologies, to train and fine-tune the underlying AI models, or advise on AI implementation projects	Developmental perspective	6
R	Educators and researchers (e.g., schoolteachers, professors) who teach AI literacy and train individuals to become familiar with AI capabilities in their daily lives	Educational perspective	5
S	Product and sales managers working for AI providers who distribute AI-based technologies to business clients; they were included to ensure that we would have a consumer-producer perspective from the business-to-business (B2B) sector as part of our analysis	Consumer-producer perspective	6

AI literacy for responsible human-AI interaction, based on the examples provided by the interviewees. By analyzing the examples and different use cases involving different AI technologies, we analyzed which components of AI literacy may be seen as increasingly relevant when transitioning from a sole focus on efficient human-AI interaction to a broader emphasis on responsible human-AI interaction. In the fourth section, we asked interviewees what factors on different levels – such as society (macro-level), their organization (meso-level), or themselves as individuals (micro-level) – influence and foster the development of AI literacy for responsibility in human-AI interaction.

3.3 Data Analysis

We followed a mainly deductive and partially inductive coding process based on Mayring (2014) for qualitative analysis of the 28 interviews. First, we aimed to identify relevant components of AI literacy that knowledge workers perceived as being relevant for responsible human-AI interaction. Following deductive coding, we used the components that we derived from the literature in the first stage (as visualized in Fig. 1) involving literacy aspects (further split into human-centred aspects and technological aspects), building blocks, etc., to establish categories with definitions, anchored examples, and coding rules (Mayring, 2014, p. 95). We then assigned text passages from the interviews to those categories and formed subsequent categories inductively during the coding process, when we explored novel phenomena that we could not aggregate within our existing categories (Mayring, 2014, p. 80). The different codes linked to the literacy aspects were then assigned to the related literacies (e.g., digital literacy, data literacy, etc.).

We also coded what hinders the development of relevant abilities and how they can be fostered, where we inductively created categories. We thereby adhered to the knowledge workers' framing of influencing factors, ensuring that our analysis remained aligned with their descriptions. We then mapped the identified influencing factors to the different analysis levels that we derived deductively in the theoretical background (see Section 2.1).

Moreover, we coded the different use cases that were addressed in the interviews, synthesized them, and categorized them into six distinct groups ("decision-making," "creation and information," "prediction," "optimization," "physical," and "communicative and supportive" tasks) to reveal any differences in skill requirements between use cases, as presented in Appendix 5. Measures to foster AI literacy were also coded inductively and linked to the specific use cases in which they were mentioned. To ensure reliability, the coding scheme was developed collaboratively by two coders. Appendix 4 provides a snapshot of the coding.

Building on this, we analyzed the coded data to answer our research question regarding what shapes AI literacy for responsible interactions of knowledge workers with AI. When analyzing the prevalence of themes in participant responses, we realized that certain components were named more often, which suggested to us a need to rethink current AI literacy frameworks. We also identified differing perspectives among knowledge workers, for istance, regarding which abilities can be developed through training and which should already be part of a person's disposition (such as moral values). Furthermore, we identified which aspects are considered essential for responsible human–AI interaction and which factors influence the development of these abilities – insights discussed further in Section 4.

3.4 Data Quality

To demonstrate the rigor and quality of our interview-based research, we provide a transparent and explicit discussion of validity. In particular, we sought to ensure descriptive and design validity by systematically conducting and carefully documenting the interviews using a structured interview guide (Sarker et al., 2018). To prevent the loss of valuable information, all interviews were transcribed in full, ensuring that the researcher's protocol was not the sole basis for analysis. In addition, each interview was conducted with two researchers present – one leading the interview and the other taking notes. This dual-researcher approach added an extra layer of interpretation accuracy beyond the transcripts alone, further strengthening validity (Myers & Newman, 2007). To enhance interpretive and analytical validity, we engaged in simultaneous transcript coding, fostering a shared understanding and consistency in interpretation. We also aimed to secure theoretical and inferential validity by systematically documenting the coding process, including defintions for each category supported by illustrative interviewee quotes, and developing a detailed coding scheme collaboratively. This scheme was continuously refined through coder discussion and repeated training, ensuring consensus and readability in the interpretation of the data (Garrison et al., 2006).

4 Results: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?

The following section presents our key findings based on the analysis of the interviews. While we do not claim to provide an exhaustive account of all components relevant to AI literacy, our aim is to contribute a perspective on how the concept of AI literacy might be rethought. We seek to move

beyond a narrow focus on efficiency and instead emphasize the need for AI literacy to support *responsible* interaction. This shift in perspective underscores the importance of equipping knowledge workers not only with the technical skills to work with AI but also with the capabilities to engage with AI systems in an ethically sound, accountable, and context-sensitive manner. We identified core components that knowledge workers perceive as enabling responsible interaction with AI, alongside influencing factors across different analysis levels that may challenge skilling processes for this enablement. While Fig. 1 presents a synthesis of existing conceptualizations of AI literacy, we focus in this results section on how the concept of AI literacy must be rethought for *responsible* human-AI interaction, resulting in Fig. 2. From the findings we suggest that while there are tensions and different views, guaranteeing *responsible* human-AI interaction requires awareness of the importance of knowledge workers' personal dispositions (Section 4.1), the need for human-centered aspects of AI literacy (Section 4.2), the need of ethical literacy (Sect. 4.3), and factors that influence AI literacy development across different levels (Section 4.4). In the following sections, we elaborate on these four key findings by summarizing the findings from our interviews.

4.1 Personal Dispositions Gain Relevance for Responsible Human-AI Interaction

Concerning the building blocks relevant to AI literacy, our data analysis outlined the importance of personal dispositions to foster individuals "feeling responsible" and thus responsibly interacting with AI technologies. Our interviews highlighted that since AI is continually evolving, AI literacy is not static but requires the motivation and mindset for continual learning and moral values to guarantee responsible interaction with AI.

Responsible AI usage was shown to go beyond having certain skills and knowledge, with a majority of codes related to what we defined as personal disposition (58.16%) compared to knowledge (25.52%), skills (11.22%), capabilities (4.08%), and competencies (1.02%). Concerning personal dispositions, patience was mentioned as important (e.g., D1, S2) as "there are a lot of stumbling blocks" (S2, p. 12) when first implementing AI, as well as forward-thinking and an "open mindset for challenges" were reported as critical (e.g., S2, M2). This means an awareness that each "[AI] project has its own difficulties, so you have to expect and embrace them" (S2, p. 12). Personal dispositions also include having a deep-seated drive for self-learning and being open to creating one's own experiences with AI (e.g., M4, D5, S5) to be prepared for rapid AI progress that leads to technical skills becoming outdated (e.g., S2, S6). At the same time, responsible interaction with AI was reported to require being able to scrutinize opposing interests with a "healthy skepticism" (D5, p.8). While a healthy skepticism does not necessarily increase performance, it was reported to be a precondition for the responsible use of AI. As one developer described: "You're not relying on it blindly, but you check it again with your own knowledge and conscience" (D1, p. 3). While this may, to some extent, contradict efficiency, it was reported as relevant to responsible

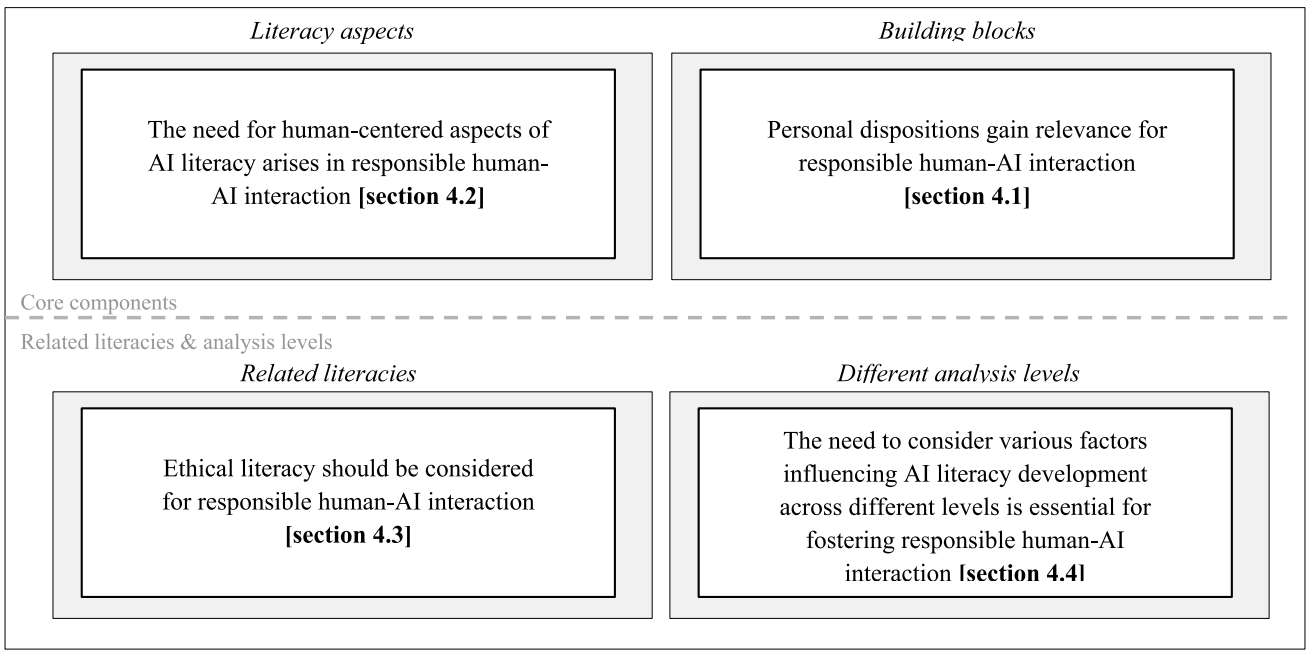


Fig. 2 Moving from efficient toward responsible AI literacy

interaction. This illustrates that personal dispositions may be highly relevant building blocks of AI literacy when it comes to responsible interaction with AI.

Yet, the analysis shows different perspectives regarding fostering personal dispositions. Some interviewees suggested that certain factors need to be anchored in a person's disposition as a prerequisite for being able to interact with AI responsibly. Others, though, argued that personal dispositions are influenced by and develop through different training. The following sections elaborate on these two perspectives.

Personal Dispositions Are a Prerequisite for Responsible Interaction With AI One perspective suggests a shift away from a competence profile involving several skills and knowledge to a profile of individuals' moral values and dispositions that make them literate to deal with AI responsibly. Some interviewees argued that personal dispositions cannot be taught or acquired through learning, and thus, individuals require a certain foundation of personal dispositions (e.g., S3, S2). As one salesperson framed it: "I believe that the skills are indeed more about the soft skills … particularly the mindset. This is because it is a skill that cannot be developed or learned but rather is intrinsically present. Only in a second step does familiarity or interaction with technology, or specifically AI, become relevant." (S2, p.9).

Moral intuition, for instance, was identified as an essential inherent personal resource to evaluate AI's output, which some expect to be anchored in a person's virtue ethics. Some suggested that since moral values differ (e.g., R1, S3), morality is lived differently, depending on the environment and shaped by experience and society. For instance, "what are the fairest criteria?" (S5, p. 13) asked one participant by stating that reflections of fairness depend on individual moral character and virtues. However, these virtues can be viewed as less learned compared to technical skills (e.g., M2). Managers who suggest that moral intuition is not really learnable reported that they would hire people based on their personal dispositions, such as personality and moral character, rather than their skills and abilities (e.g., M2, M6), meaning those managers prioritize the hiring of individuals with desirable virtues over those with specific skills, believing that certain personal dispositions are integral for responsible decision-making and behavior.

Individual Personal Dispositions Are Influenced by and Developed Through Increased Human-AI Interaction Some knowledge workers suggested that the personal dispositions of individuals are influenced through their interaction with AI technologies. For instance, in the context of recruiting professionals, AI could foster objectivity in knowledge workers' decision-making (H3). This is because AI's neutrality in evaluating each situation based on equal criteria could rub off on them, influencing their moral assumptions and moral values: As one HR professional stated, "As the evaluation of applicants gets more objective through AI, you also start to avoid looking at the photos of applicants to make a decision and instead check only qualifications. I think that, in this respect, more objectivity through AI is a good thing" (H3, p. 10). Likewise, some argued that "most people question [AI]" (M5, p. 5) or "check [AI's] outputs by doing research" (D5, p. 6–7), which can develop their meta-skills when reflecting on "What should I do with data? What should I instead not do with it?" (D5, p. 10).

While both perspectives are legitimate, in the end, an individual's personal disposition and training may be mutually enriching. One salesperson's statement illustrates this: "A mindset, you cannot simply train or learn it – it is instead something intrinsic, maybe arising when you are getting familiar with specific AI technologies and increasing your affinity … through learning by doing" (S2, p. 9).

For personal disposition to develop, organizations aim to encourage a learning culture of errors, implying that the teams share responsibility (M6, S2). As one interviewee suggested what is a essential resource for individuals in organizations working with AI responsibly: "great tolerance of failing … to fail often but learn fast" (S2, p. 16). By promoting a learning culture that embraces errors, knowledge workers are encouraged to take ownership of mistakes without fear of negative consequences from management (e.g., M6, S2). It can also promote ethical dialogue among different professionals (R2) around what is considered right or wrong in various tasks, encouraging reflection and discussion on the underlying reasoning (S2).

4.2 The Need for Human-Centered Aspects of AI Literacy Arises

It is notable that our interviewees, having provided different use cases, had different perspectives concerning which aspects would become relevant in human-AI interactions. Some highlighted that socio-cultural aspects (41.64%), such as communicating and creative thinking, rather than technical aspects, are perceived as becoming increasingly relevant for responsible human-AI interaction, especially when considering user communication through and with generative AI applications. In contrast, developers highlighted the increased complexity of AI models and stressed the importance of developmental (63.64%), critical (33.33%), and functional aspects (31.25%) for those responsible for deployment of AI technologies within business and society, implying the need for a more profound technical understanding on

the part of leaders, managers, and executives who drive AI implementation strategies. Likewise, educators focused on functional (31.25%) and physical (33.33%) aspects to better integrate AI and robotics into existing curricula. This shows that different stakeholders hold different views, depending on the AI artifact, with respect to what is essential for responsibly interacting with AI, impacting their "acting responsibly," which we elaborate on in the following.

AI's Increased Complexity Requires a More Profound Technical Understanding AI was reported to gain increased autonomy and be used for tasks that can have an impact on professionals. That is why our interviews suggest that individuals should have objectives that go beyond blindly following AI and should reflect upon the input they provide AI, which requires a certain level of technical understanding. For instance, "Chat-GPT [has] a relatively simple and generic interface, [and so] the competence to assess or understand that not everything the thing says is correct [is a] minimum requirement" (D1, p. 4). This becomes particularly relevant when interacting with smart devices, in which AI's activities are not directly linked to prompts made by the human user (D1).

One salesperson explained what would be considered a minimum requirement for responsibly interacting with AI: "I should have a high level of competence in the field of statistics … The thing is, you probably needed that already a few years ago when you were dealing with [AI models]. Today, I can just somehow use AI tools out of the box, acquire them somewhere within an overall software package, and not even really know what's happening in the background. So, depending on the use case, it would actually be necessary to take a close look at how the AI works behind the scenes." (S4, p. 11). The salesperson added an example about assessing people's creditworthiness: "It's no longer just about creditworthiness in six categories, like it was maybe decades ago. Now, when assessing creditworthiness, I believe 80 or 90 characteristics are easily taken into account. It would be preferable that if I rely on this score, I agree with the criteria being used" (S4, p. 11). This shows that increased accessibility and usability of AI technologies may lead to easy usage of such assistance, but awareness of how the code works is also required when aiming to deal with AI responsibly.

AI's Increased Usability Requires Less Technical Understanding but Increased Domain Knowledge and Communication Skills While the growing complexity of AI necessitates a more profound technical understanding, the enhanced usability of generative AI technologies was also reported to demand fewer technical aspects from users, with the focus instead on social and creative skills – depending on the use case and type of AI.

In the context of using chatbots, for instance, some interviewees suggested that collaborating with AI responsibly requires increased communication skills (e.g., S2, D3). Users are required to utilize inclusive language to responsibly influence the output of AI tools (H3). As one interviewee explained, "I only have an influence on what comes out through the words I use. And depending on how precise I am in my description, I have different [outputs], and it's up to me to help shape them" (H3, p.1). This is why AI may even require improvement of communication skills, as one developer argued: "Making perfect prompts becomes more relevant, so what criteria do we need for the prompts that we put in there to get the quality that we need?" (M6, p. 3–4). Another interviewee supported this, stating, "[When] I interact with an AI-supported system, I should also just allow myself fewer spelling mistakes for the sake of simplicity, for example, or to get to the point more quickly" (S2, p. 6). A salesperson experienced an enhancement of communication skills, reporting "I use different generative AI tools [and] I just notice [that] there is also a certain learning effect for me now. Depending on what information I feed a particular tool, the answers are reasonably useful" (S4, p. 5).

One example of the increased relevance for domain knowledge was provided by a salesperson from a startup that has designed an app that analyzes images of dogs to predict their mood and needs, offering actionable suggestions for owners such as "you should go for a walk" (S6, p. 7) The app is intended to address the knowledge gap of some dog owners, as "there are indeed people who are completely unable to recognize how the animal is feeling" (S6, p.7). The salesperson emphasized that the app is designed to be as user-friendly as possible, requiring no technical skills or understanding of the model. However, the salesperson acknowledged the potential risks, noting that if the app says "the dog is fine, you don't need to go to the vet, and then the dog dies – that would, of course, be truly terrible" (S6, p.8). This highlights the ongoing need for user domain knowledge to guarantee responsible interactions, even if the AI is designed to compensate for a lack of domain knowledge.

Overall, we conclude that AI's increasing complexity necessitates technical skills to comprehend and manage the sophisticated algorithms and data processes that drive its functionality, particularly on the part of developers. Conversely, the growing usability of generative AI in particular has significantly reduced the need for technical skills, allowing more knowledge workers to use AI tools with minimal understanding of the underlying mechanisms – often requiring only basic prompting knowledge. This shift, however, places greater emphasis on domain knowledge: in order

to use generative AI responsibly and purposefully (i.e., to avoid misleading or irrelevant results), users must be able to formulate inputs as concretely and contextually as possible, and critically reflect on and adapt the output to their specific subject matter. For example, a marketing professional using a generative AI tool to draft campaign slogans does not need to understand the model's architecture, but must be able to judge whether the suggestions align with brand tone, target audience, and compliance guidelines (M6).

Combining the findings, we suggest that balancing technical aspects with domain knowledge in the sense of a critical mindset and strong communication capabilities is essential to interact with AI responsibly and, thus, to foster the "acting responsibly" of individuals. However, our findings also indicate that not all AI literacy aspects apply equally to every professional role, as each use case demands a distinct combination of competencies. This emphasizes the importance of context and the type of AI in shaping AI literacy requirements. Also, given that required skills vary significantly across roles – for example, between developers and managers – role- and domain- specific training initiatives are viewed as both necessary and effective (e.g., M6, H3). This further emphasizes the need for knowledge workers to recognize their shared responsibility, particularly in contexts where their skill levels differ.

4.3 Ethical Literacy Should be Considered

Our interviews reveal that AI literacy is particularly influenced by ethical aspects that were reported as critical for fostering responsible human-AI interaction. Notably, most of our codes about abilities required for responsible human-AI interaction can be linked to ethical aspects (45.16%), while fewer pertain to data literacy (27.42%), algorithmic literacy (9.68%), digital literacy (6.45%), and others (11.29%). Based on these findings, we argue that having only certain abilities linked to ethical aspects, such as critical reflection, would not be enough to guarantee responsible interaction with AI. Rather, we would argue from the interview data that we require *ethical literacy*, which involves reflecting various ethical viewpoints, understanding underlying ethical principles, and recognizing the moral values of different communities.

For instance, our interviews highlighted the scenario of an AI-driven recruitment system used to screen job applicants (e.g., H1, H5). As an example, one interview (H1) suggests that an AI specialist with strong AI literacy may possess a detailed understanding of how an algorithm processes applications – how it weighs different variables, identifies patterns, and makes recommendations. Further, the specialist might also recognize the potential biases embedded in the system due to historical inequalities in the training data (H1), which reflects an awareness that falls under the ethical aspects of AI literacy. Both are required, however, since technical expertise without knowledge about ethical principles alone is insufficient (M6). Without ethical literacy, the specialist may fail to grasp the broader societal implications of these biases (R1) or critically reflect on whether automating hiring decisions aligns with fundamental principles of fairness and human dignity (M2). They might acknowledge that certain groups are disadvantaged, but there remains a risk that they rationalize this as an acceptable trade-off for efficiency, overlooking human rights concerns on a more systemic, meta level (D3). Moreover, knowledge workers need to recognize that moral values and ethical expectations surrounding AI applications differ across societies, as an AI-driven hiring system deemed appropriate in one cultural context may be considered deeply unethical in another (D3).

Therefore, people who may be highly AI literate but lack ethical literacy can also use AI irresponsibly. For instance, knowledge workers who can exploit system loopholes for personal gain could be considered highly AI literate in a technical sense, but their lack of ethical awareness of potential harm to others and their non-compliance with given laws and norms (linked to feeling responsible) undermine responsible human-AI interaction. Thus, we conclude that ethical literacy is crucial for fostering responsible human-AI interaction, as it contributes to "being responsible" by providing knowledge about ethical assessments and roles. Based on our interviews, we can derive two different perspectives on how ethical literacy intersects with AI literacy, discussed below.

Ethical Literacy can be Perceived as a Prerequisite for AI Literacy One perspective that emerged in the interviews suggests that ethical literacy is foundational for AI literacy. Knowledge workers were reported to need a profound understanding of ethical behavior before engaging with AI technologies. For instance, it was suggested that AI developers require ethical literacy to avoid ethical dilemmas in a system's design and to build norm-compliant systems, as programming skills alone, without ethical literacy, would not guarantee responsible AI development (M6). As "building such systems requires certain skills to address ethical issues within the system's architecture well before they manifest during user interactions" (M6, p. 15). One developer suggested that ethical considerations become particularly relevant when AI's output impacts knowledge workers: "Ethical knowledge is crucial for those responsible for AI system outputs that make predictions affecting human lives" (D5, p. 10). Concerning this, one manager recommended that "organizations should define their own

understanding of ethics in general before considering AI ethics" M6, p. 12).

Although those examples could imply that ethical literacy is necessary primarily for AI designers, our interviews indicate that user ethical literacy is equally important to prevent misuse. One interviewee, in this context, pointed to how AI can facilitate breaking into IT systems, "as AI makes it easier for someone to be a hacker with little experience" (D6, p. 6) compared to previous technologies. This shows, while AI lowers the barriers to entry for interacting with technology, it may simultaneously increase the risk of unethical use when lacking ethical considerations. This example further illustrates that, in certain cases, AI literacy alone (effectively using the AI system) may be insufficient for ensuring the responsible use of AI technologies. "A basic understanding of ethical values like equality, inclusion" (S5, p. 8) was reported to be required to mitigate the risk that users manipulate the technology for personal gain, potentially leading to misuse if ethical literacy is lacking (S5). Absent ethical literacy, knowledge workers who exploit AI loopholes for personal gain and do not feel responsible for the outcomes fail to recognize that such actions can harm others and have legal ramifications (S3, p. 18). In this regard, ethical literacy can serve as a safeguard against misuse and harm to others by clarifying rules and consequences before the interaction. From this perspective, ethical literacy can be seen as a prerequisite to AI literacy.

AI Literacy can be Perceived as a Prerequisite for Ethical Literacy Some interviewees suggested that a certain level of AI literacy is necessary to recognize and address ethical issues. Understanding the complexities and underlying mechanisms of AI technologies was reported as crucial for individuals to responsibly manage AI's unintended consequences (e.g., M4, H1) and, thereby, develop their ethical literacy. Without a foundational level of AI literacy, recognizing responsibility issues was perceived as challenging (e.g., R1, R3, R4).

Particularly when decisions are based on AI recommendations, we suggest that decision-makers must first understand AI mechanisms, which then informs their ethical literacy in making responsible decisions. As one interviewee stated, "An AI can't recognize a person's value … It might suggest someone for a management position, but this person could mistreat employees, as AI cannot assess a narcissist's behavior. AI might use facial recognition to analyze word choice, but it cannot conduct a full assessment" (M6, p. 10). Being aware of AI's functioning can, thus, be seen as a foundation for knowledge workers' ethical literacy, which is then required for deciding whether the context is applicable for using AI. As another interviewer said, "I would not hire a receptionist through AI because I hire based on charisma and personality, which makes knowledge workers feel comfortable. This might differ when hiring a software developer, where the AI could be more suitable for assessing programming skills" (H1, p. 8).

These two perspectives illustrate that ethical literacy is essential to responsibly develop and interact with AI systems, and recognizing ethical issues requires a basic understanding of AI's functioning (linked to AI literacy), highlighting the need for multi-literacy. More precisely, some organizations view ethical awareness as a prerequisite for the responsible use of AI, while others emphasize the importance of developing ethical competence through experiential learning, particularly once foundational technical skills, such as programming, are in place. Since AI progress is rapidly evolving, the findings combined suggest that achieving responsible human-AI interaction requires that knowledge workers, organizations, and society have a certain baseline of moral values that are more important than technical knowledge: "You don't necessarily need a deep technical understanding of AI … rather, you need an understanding of the impact that AI can have" (R2, p 8). This implies, while technical knowledge in human-AI interaction can be acquired and refined through continuous learning, individuals must possess certain personal dispositions as moral values that underpin and promote responsible engagement with AI.

4.4 Various Factors Influence AI Literacy Development Across Different Analysis Levels

Our interviews indicate that several factors influence whether and how AI literacy can be fostered to ensure responsible human-AI interaction. While those influencing factors should not be considered as components of AI literacy, they need to be considered when conceptualizing AI literacy to answer the question of what shapes AI literacy for responsible interactions of knowledge workers with AI. As illustrated in the theoretical background (see Section 2.1), AI literacy can be analyzed on several levels. In our empirical data, however, we identified influencing factors at the individual level (micro-level, 47.47%), organizational level (meso-level, 37.74%), and societal level (macro-level, 14.79%) that influence how knowledge workers' AI literacy can develop and thus needs to be considered when holistically conceptualizing AI literacy for responsible human-AI interaction. These influencing factors could enhance AI literacy by facilitating or fostering upskilling mechanisms or diminish it through deskilling mechanisms when interacting with AI. Interestingly, most factors influencing deskilling reported were situated at the micro-level, whereas most

factors that were suggested to influence upskilling were identified at the meso- or macro-level, which we explain in the following.

Deskilling is Often Influenced by Factors at the Micro-Level Our analysis highlights that increased usability, easy access, and human-like and intuitive design of AI technologies can decrease the need for AI literacy for effective interaction. Several interviewees noted that when AI systems are easy to use and understand, knowledge workers can feel less need to develop or apply certain AI literacy abilities (e.g., D6, R4, M4, S2). Specifically, AI's personalized and human-like communication, coupled with its immediate output in natural human language, can lead to a tendency to "unlearn how to question things" (D2, p. 6). However, this may undermine a responsible human-AI interaction when unintended consequences are considered. While human-like AI may not free individuals from being responsible for outcomes, it may lead knowledge workers to act less responsibly in their interactions with AI. As one interviewee observed, this may lead to excuses such as "we can't do anything about it; the system said so …" (H1, p. 10). The opacity of data processing behind AI outputs may exacerbate this issue, making critical reflection difficult (e.g., R5) and challenging to act responsibly through deskilling.

This issue was particularly evident in our interviews when it came to using generative AI applications, in which knowledge workers can interact with AI easily using natural language (e.g., M1, S2, D3). While social aspects of AI literacy – such as communication and critical aspects in evaluating outputs – become increasingly significant in this context, individuals may unlearn writing creative texts on their own (e.g., D1). As one knowledge worker explained: "If every company now starts using ChatGPT to contact applicants or customers, then of course the variety in how it's done will likely be rather limited. That means you no longer stand out" (D1, p. 10). In contrast, for instance, discussion about AI-based drones for error detection emphasized the importance of technical aspects for correct operation (M5). This shows that the type of AI technology may serve as a critical influencing factor that impacts AI literacy development at the individual level.

Another deskilling moderator at the micro-level could be the trust users place in AI technologies based on misleading perceptions of AI (R1). One researcher noted that "many people anthropomorphize AI systems and overestimate them," leading to overtrust in AI's capabilities and the mistaken belief that AI can perform tasks as effectively as knowledge workers (e.g., R1). This overtrust can inhibit the development of critical abilities to challenge AI outcomes. As one manager warned, "I see a big danger that people will question results less because of the easy way of using generative AI" (M1, p. 6). Thus, we suggest that overtrust in AI and the high usability and human-likeness of AI technologies, both micro-level moderators, can lead to deskilling and thus potentially reduce AI literacy and negatively impact responsible human-AI interaction.

Upskilling is Often Influenced by Factors at the Meso- and Macro-Levels Our interviews revealed culture-dependent factors at the meso-level that can lead to upskilling, enhancing AI literacy for responsible human-AI interaction. For instance, one manager noted, "The more technical an organizational unit, the more critically employees view AI compared to those working in marketing, who approach AI more playfully or hedonistically" (M3, p. 5). This suggests that department culture may influence whether AI technologies are applied and, consequently, whether new skills can be acquired through interacting with AI. A culture of distrust in data and novel technologies, in contrast, can hinder upskilling within organizations (M5) – except when individuals already have high AI literacy and independently use AI in opposition to the organizational culture. This was illustrated by one manager, who stated, "I analyzed data at home to get insights but presented the findings at work as if they were based on gut feeling to get things through" (M5, p. 3). As for overcoming such distrust, one manager suggested that fostering a critical and open-minded work culture is essential for the responsible deployment of AI within organizations (e.g., M2), particularly given the work culture's impact at the individual level (micro-level).

At the macro-level, our interviews further highlighted that societal moral values (e.g., R2), national resources for AI innovation (e.g., R3), and the level of AI maturity within a country (e.g., D3) may serve as key factors moderating upskilling that enhance AI literacy for responsible human-AI interaction. Specifically, adult education (D1, M6) and a sense of social duty (R2) were identified as important. Regulatory frameworks at the macro level (such as the AI Act for AI regulation) also define responsibilities (e.g., S3, D7, R1) and therefore shape "being responsible" in human-AI interactions. One developer emphasized the need to leverage AI opportunities to promote inclusivity: "We have to use AI's opportunities to include people who didn't have the opportunity before, especially [through] text-to-speech or speech-to-text, so that all individuals can interact with others" (D3, p. 7).

Organizations recognize the need to critically reflect on potential deskilling that could undermine responsible human-AI interaction and understand that addressing these influencing factors across different levels is essential for fostering the responsible use of AI technologies (e.g., M1, H1).

To address this, some propose conducting targeted audits to identify which competencies remain essential for responsible AI use and which may become less relevant (e.g., R2), for instance, through AI's increased usability (D1). Given that required skills vary significantly across roles – for example, between developers and managers – role-specific training initiatives are viewed as both necessary and effective (e.g., M6, H3). Also, several interviewees further suggested that gaps in specific AI literacy dimensions – such as domain knowledge – can be mitigated through collaboration with subject-matter experts. As one salesperson explained, that is why "we enter a co-creation phase together with the client" (S2, p. 6).

4.5 Rethinking the Concept of AI Literacy

As visualized in Fig. 2, our findings suggest that the concept of AI literacy requires reconsideration when aiming for responsible human-AI interaction, particularly in light of four key insights. Figure 3 provides further detail, based on the descriptive results of our interviews, illustrating that certain AI components gain in relevance and certain factors need to be considered when shifting from only interacting with AI efficiently to human-AI interaction that is truly responsible. The interviews thus help us to better understand what shapes AI literacy for (ir)responsible interactions of knowledge workers with AI and provide valuable insights for AI literacy conceptualizations.

First, concerning *building blocks*, we suggest from the findings that personal dispositions are foundational to responsible human-AI interaction, as they shape how knowledge workers perceive and assume responsibility for outcomes involving AI. Responsibility gaps can emerge when such dispositions are lacking, potentially undermining responsible decision-making in AI-supported environments. These dispositions are not static; they evolve through ongoing interaction with AI systems, highlighting the dynamic nature of human-AI engagement. Regarding *literacy aspects*, there is a growing need – as AI becomes more powerful – for human-centered aspects to ensure responsible use and oversight. Simultaneously, the increasing usability of AI systems reduces the need for deep technical expertise but raises the importance of domain-specific knowledge and communication skills. With respect to *related literacies*, ethical literacy emerges as a critical component of responsible AI interaction and is often viewed as a prerequisite for effective AI literacy. Conversely, acquiring AI literacy can also serve as a foundation for developing ethical reasoning in AI contexts, suggesting a reciprocal relationship. The combination of technological and non-technological literacies – particularly digital, data, and ethical – has a stronger impact on responsible AI use than any single literacy

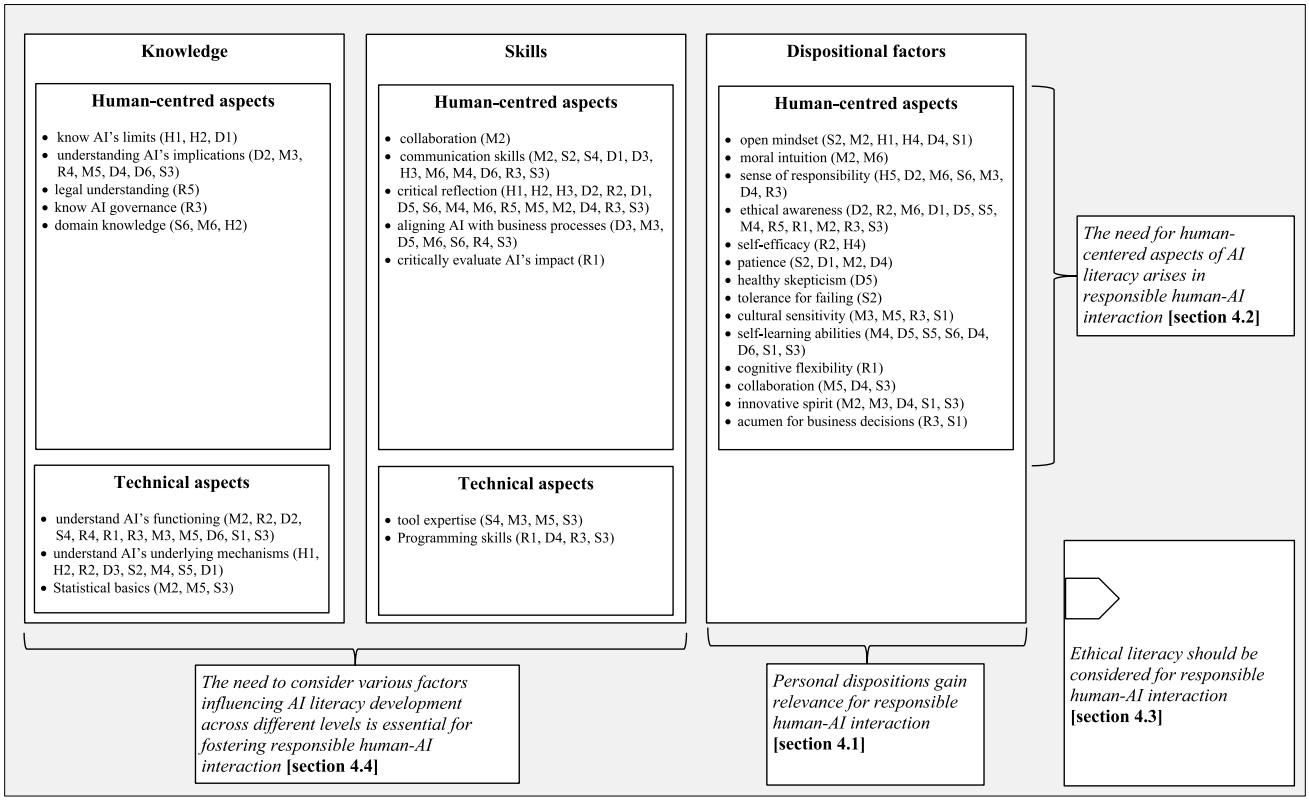


Fig. 3 Connections between the data and the four key findings

alone. When considering *different analysis levels* of AI literacy, the phenomenon of deskilling, particularly driven by micro-level factors such as overreliance on user-friendly AI systems, underscores the importance of addressing skill retention and development across individual, organizational, and systemic levels.

5 Discussion

In this section, we present the theoretical and practical implications derived from our study findings. Table 2 is an overview synthesizing the key insights discussed above, serving as the foundation for our propositions and the corresponding recommendations.

As discussed in our theoretical background, we assert that the assignment of responsibility is legitimized through two interconnected dimensions: individuals' ability to navigate specific situations (linked to *acting responsibly*) and the expectations tied to their professional roles (linked to *being responsible*). We argue that alignment between abilities and roles, what we term the role–ability fit, is essential for fostering genuinely responsible collaboration between humans and AI systems. Neglecting this alignment risks misunderstandings, inefficiencies, and even irresponsible AI use. However, our interviews challenge this assumption by revealing that such misalignment is, in fact, common in practice.

To avoid misalignment and inform organizational management, we discuss how certain problems may challenge responsible human-AI interactions with respect to *being* responsible and *acting* responsibly with AI, and suggest four future research propositions that need further investigations (Section 5.1). With respect to the practical implications, we provide recommendations for organizational strategies that may help manage human-AI interaction more responsibly (Section 5.2). We end the discussion with limitations (Section 5.3).

Table 2 Synthesis of key findings, derived propositions, and practical recommendations

Identified problems	Findings and tensions [Section 4]		Propositions [Section 5.1]	Recommendations [Section 5.2]
The *responsibility problem*: Being responsible requires an understanding of shared responsibility	Personal dispositions gain relevance for responsible human-AI interaction [Section 4.1]	Personal dispositions are a prerequisite for responsible interaction with AI Individuals' personal dispositions are influenced by and developed through human-AI interaction	Knowledge workers' personal dispositions influence how responsible they feel for outcomes of their interactions with AI; lacking certain dispositions can lead to responsibility gaps	Establish a common sense of shared responsibility among knowledge workers, each having individual personal dispositions, by fostering ethical dialogue on the right and wrong uses of AI in different situations and for different tasks and discuss the reasons
	The need for human-centered aspects arises in responsible human-AI interaction [Section 4.2]	AI's increased complexity requires a more profound technical understanding AI's increased usability requires less technical understanding but increased domain knowledge and communication skills	The interplay between technical aspects and domain-specific knowledge influences responsible human-AI interaction, with varying importance depending on AI systems and task context; this challenges the alignment of responsibilities with individual capabilities	Conceptualize AI literacy as a "toolbox" and understand it as a flexible framework that can be adapted to various human-AI interaction contexts fostering a sense of shared responsibility
The *skilling problem*: Acting responsibly requires an understanding of multi-literacies	Ethical literacy should be considered for responsible human-AI interaction [Section 4.3]	Ethical literacy can be perceived as a prerequisite of AI literacy AI literacy can be perceived as a prerequisite for ethical literacy	The integrated development of technological and non-technological literacies exerts a greater influence on fostering responsible human-AI interaction than the advancement of any individual literacy in isolation	Recruit for moral values, drive for continuous learning, and focus on organizational training that aims to strengthen these, rather than training only in AI-related knowledge and hard skills
	The need to consider various factors influencing AI literacy development across different levels is essential for fostering responsible human-AI interaction [Section 4.4]	Deskilling is often influenced by factors at the micro-level Upskilling is often influenced by factors at the meso- and macro-levels	AI's increased user-friendliness leads to deskilling, with individuals relying less on their own abilities, thus diminishing the skills needed for acting responsibly in human-AI interaction	Implement targeted and role-specific AI literacy training by assessing the most critical abilities required for different professional roles and fostering learning from one another

5.1 Theoretical Contributions: How Certain Problems may Challenge Responsible Human-AI Interaction

This section discusses problems derived from our findings that can challenge responsible human-AI interaction. We discuss these findings in relation to our theoretical background, proposing that responsible human-AI interaction requires alignment between *being responsible* and *acting responsibly* to guarantee what we framed as the *role-ability fit.* We address problems related to knowledge workers' responsibilities in human-AI interactions (*being responsible*) and those related to knowledge workers' ability to act responsibly in these interactions (*acting responsibly*). As illustrated below through the four propositions, those problems may influence the management of human-AI interaction, for which we provide managerial guidance as practical implications in Sect. 5.2.

5.1.1 The Responsibility Problems: Being Responsible Requires an Understanding of Shared Responsibility

While *being* responsible depends on accountabilities assigned per duty and defined per role (Hamilton & Hagiwara, 1992), there can be certain tensions in practice, as our interviews suggest, and whether individuals assume responsibility depends significantly on their personal dispositions [see Section 4.1]. Human-AI interaction involves multiple knowledge workers collaborating directly or indirectly with AI, with points of contact before, during, and after AI usage (Heyder et al., 2023). Given this complexity, our empirical findings highlight a greater emphasis – compared to existing AI literacy frameworks – on personal dispositions such as moral values. These dispositions appear to play a critical role in fostering knowledge workers' willingness to assume responsibility, particularly in situations where lines of accountability are blurred – for example, when no individual is formally designated as responsible or when individuals do not feel accountable for the outcomes of human-AI interactions, an occurrence similarly noted in previous studies (Mayer et al., 2020; Passlack et al., 2024).

Accordingly, our findings address the prevailing ambiguities surrounding responsibility among the various knowledge workers involved in the development, deployment, use, management, or education of AI systems. While it is argued that AI should be entrusted with final decision-making, given its perceived potential to be less error-prone (e.g., S3), our findings also reveal concerns that knowledge workers may become less inclined to assume responsibility when interacting with AI systems – particularly when they lack the AI literacy necessary to fully understand how such systems function (e.g., D4, H1). Some argue that even developers no longer understand the algorithms, as AI models' complexity may lead to greater opacity concerning how the system generates results and can lead to blind following by users (e.g., H1). The results further indicate a diffusion of responsibility: for instance, developers may view users as responsible for ensuring ethical interaction with AI, whereas users often assume that the AI itself will operate ethically. Thus, future studies would be required to explore how personal dispositions such as patience and moral values can be systematically integrated into AI literacy training and education. Research could examine how these character traits can be cultivated and measured in various organizational settings and how they contribute to responsible AI implementation and management.

We propose:

> **Proposition 1**: *Knowledge workers' personal dispositions influence how responsible they feel for outcomes of their interactions with AI, while lacking certain dispositions can lead to responsibility gaps*.

As AI becomes more advanced, powerful (Berente et al., 2021), and capable of causing more damage if misused (e.g., M3), some knowledge workers perceive that users bear a greater degree of responsibility for its use compared to earlier, less-sophisticated technologies (e.g., D1, S4). In other words, the advanced capabilities of AI require a heightened awareness and accountability on the part of those who utilize it. While some argue that more complex AI models necessitate deeper technical understanding (e.g., R1, D4), others suggest that increased usability reduces the need for technical expertise, shifting the emphasis instead toward domain knowledge and communication skills [see Section 4.2]. A lack of technical understanding can hinder individuals' ability to comprehend the underlying code, which may be essential for making informed decisions relevant to responsible AI use. Conversely, possessing technical expertise without sufficient domain knowledge limits one's ability to assess whether a model's functioning – though technically understood – is appropriate within a specific context where domain-specific insight is crucial (e.g., S6).

In addition, when roles are ill-defined or responsibilities are ambiguously distributed, it can lead to uncertainties and misalignment between required and actual capabilities, hindering an appropriate role-ability fit. For instance, our interviews lead us to suggest that everyone bears a certain level of responsibility and that different knowledge workers are involved in the process of improving AI technologies and have ethical concerns (e.g., S3, M6), given the subjective nature of ethics and moral assumptions. Further studies are needed to better understand how combinations of technical proficiency and domain-specific knowledge contribute

to responsible AI interaction. Research could examine the conditions under which either type of expertise becomes more critical, and how interdisciplinary skill sets affect individuals' ability to detect, interpret, and respond to ethical challenges arising from AI deployment. Otherwise, lines of responsibility can be blurred (Giermindl et al., 2022), thereby undermining individuals' confidence in their role expectations. We derive the following proposition:

> **Proposition 2**: *The interplay between technical aspects and domain-specific knowledge influences responsible human-AI interaction, with varying importance depending on AI systems and task context; this challenges the alignment of responsibilities with individual capabilities*.

5.1.2 The Skilling Problems: Acting Responsibly Requires an Understanding of Multi-Literacies

Acting responsibly in human-AI interaction presupposes the necessary capabilities, yet there can be certain tensions in practice flowing from whether individual skills are enough or require deeper knowledge. Concerning this, some findings indicate that possessing individual ethical skills, such as critical reflection, may be insufficient on its own to ensure responsible use. In contrast, we derive from the findings that ethical literacy emerges as an essential prerequisite for responsible human-AI interaction, as AI literacy alone would not account for responsible interactions with AI [see Section 4.3]. We propose that ethical literacy equips individuals with an understanding of human rights and the ability to navigate conflicting moral values when unintended negative consequences arise. It is essential to prevent inequalities that result from norm-compliant behavior, such as through the misuse of AI or discrimination through data biases (e.g., M3, D6). Consequently, we conclude that ethical literacy and AI literacy are dependent, and both are required for responsible human-AI interaction.

This leads us to suggest that responsible human-AI interaction must focus on the multi-literacies required for enabling individuals to act responsibly. As an example, while a developer may possess a high level of AI literacy with respect to its functional aspects, the concept of multi-literacy suggests that developers would also need ethical literacy to develop AI responsibly and inform their AI literacy. Ethical literacy is about being aware that there are different normative positions with respect to how ethics and morality can be interpreted in general. We differentiate it from the ethical aspects of AI literacy, which pertain specifically to the ability to recognize and navigate ethical conflicts within AI interactions, such as understanding issues of bias, fairness, and transparency in algorithmic decision-making, as suggested by prior research (Ng et al., 2021). While further empirical investigation is necessary to explore these connections in depth, the data generally support the importance of a multi-literacy approach in achieving responsible human-AI interaction. Thus, research is needed to examine the role of non-technological literacies within AI-related contexts, particularly ethical literacy, by reflecting on the concept of multi-literacy, with different literacies influencing each other mutually. We propose:

> **Proposition 3**: *The integrated development of technological and non-technological literacies exerts a greater influence on fostering responsible human-AI interaction than the advancement of any individual literacy in isolation*.

Other challenges regarding skill development became evident when it comes to *acting* responsibly in human-AI interaction. While there was a consensus that AI increases the number of tasks that knowledge workers are capable of and thus upskills them (e.g., S3, M6, M3, S2), practitioners observed situations that suggest knowledge workers lose certain skills, which may negatively affect their ability to act responsibly (e.g., M1, R1, D2). The problem, therefore, is to determine the main influences on human-AI interaction, leading to upskilling or deskilling that may negatively affect knowledge workers' ability to act responsibly.

Our findings emphasize the importance of examining the interdependencies between various influencing factors at individual, organizational, industrial, and societal levels as the alignment – or lack thereof – of individual values, technological factors, and cultural norms across these levels can impact the skilling of the workforce, thus, responsible human-AI interactions [see Section 4.4]. Therefore, misalignment among these factors can result in irresponsible action, while alignment can promote responsible behavior associated with *acting responsibly*. As the results show, some interviewees suggested upskilling, in that individuals increasingly reflect on decisions and critically reflect on their own opinions when interacting with AI (e.g., M3). Likewise, the research shows that the use of AI may increase knowledge workers' critical reflection skills when analyzing AI's output (Hemment et al., 2023; Passlack et al., 2025). This results from seeing AI as a second and less-subjective opinion. However, this would require that AI has no biases that can foster unintended, irresponsible outcomes. In contrast, interviewees suggested that increased ease of use of AI systems means fewer skills are required to interact with AI, which can, in turn, deskill knowledge workers (e.g., M1, R1, D2) and impede responsible acting. The accessibility and easy usability of AI may sometimes demotivate users from trying to investigate how an AI works, since little

understanding is required to interact with that AI (e.g., D6, R4). This can be aggravated by the fact that over-trust in AI can hinder critical reflection, given that users expect AI to have a high degree of accuracy and correctness (e.g., S3). Thus, future research could focus more on the interdependencies between different factors that may influence knowledge workers' upskilling and deskilling, such as conducting a cross-cultural comparison. Such investigations could also consider AI technologies' explainability and user-friendliness, or explore how users' values influence their engagement with AI artifacts differently (e.g., in a responsible or irresponsible manner) in different cultures. We propose:

> **Proposition 4**: *AI's increased user-friendliness leads to deskilling, with individuals relying less on their own abilities, thus diminishing the skills needed for acting responsibly in human-AI interaction*.

5.2 Practical Implications: How Organizations may Strategically Foster Responsible Human-AI Interaction

The interview findings further reveal practical implications with respect to how organizations may address the problems that can challenge fostering responsible human-AI interaction. We elaborate on these practical implications in the following and provide four recommendations for organizations that aim to introduce human-AI interaction effectively and responsibly.

5.2.1 Addressing the Responsibility Problems

Concerning the *responsibility* problem, our findings highlight that knowledge workers require a certain level of morality to address the responsibility gap. Given that human-AI interaction can have implications beyond fulfilling a certain task, several knowledge workers' literacies that influence human-AI interaction or that are influenced by its consequences must be considered, as also suggested by Carolus et al. (2023). Bringing diverse and interdisciplinary knowledge workers together when deciding upon AI strategies (as suggested by e.g., R2) in organizations is also required to make decisions that are aligned with ethical ideologies and moral values, which can differ between individuals and thus affect their sense of shared responsibility.

While developers should not rely solely on user moral values to ensure the responsible use of their AI technologies, users indeed bear a certain responsibility for the outcomes that result from their interactions with AI. AI interaction involves multiple stakeholders who collaborate with AI at different points, such as developers designing AI, managers implementing AI, or policymakers regulating AI, suggesting a process view of AI literacy. This is better illustrated by literature suggesting that human-AI teams "are together embedded in an encompassing organizational system, with boundaries and linkages to the broader system context and task environment" (Kozlowski & Ilgen, 2006, p. 79).

Each stakeholder requires different but complementary literacy components. Thus, dialogue among different knowledge workers and departments is required (e.g., R2), also supported by literature suggesting critically examining and clarifying shared roles and responsibilities at all levels in AI governance (e.g., Cañas, 2022). Consequently, we recommend:

> **Recommendation 1**: *Establish a common sense of shared responsibility among knowledge workers, each having individual personal dispositions, by fostering ethical dialogue on the right and wrong uses of AI in different situations and for different tasks and discussing the reasons*.

Moreover, as the interviews revealed, responsible human-AI interactions are influenced by diverse factors and are contingent on knowledge workers' use cases, making AI literacy a dynamic concept and leading to uncertainties in skill requirements. Thus, understanding AI literacy as a "toolbox" provides a flexible framework that can be adapted to various human-AI interaction contexts. We propose that understanding AI literacy as a toolbox implies that each use case requires a unique combination of specific aspects and building blocks, and is related to other literacies and influenced by factors across different analysis levels. There can be certain interdependencies between components; for instance, only the combination of ethical aspects and development aspects would account for a responsible human-AI interaction. For instance, ethical literacy and critical reflection on AI's impact should be prioritized in high-stakes decision-making tasks such as medical AI applications or legal AI advisors (e.g., R1). Conversely, other literacies gain greater importance in creative AI applications, such as generative AI for content creation, which are more closely tied to social interaction and communication (e.g., D1, H3). Theoretical models of AI literacy should, therefore, integrate these dynamic combinations rather than treat AI literacy as a static, one-size-fits-all construct.

This aligns with previous quantitative investigations on AI literacy in which measurement items of AI literacy are closely linked to the investigation context and the underlying technology being used (e.g., Kandlhofer et al., 2016; Lin et al., 2020; Ng et al., 2021). General scales for AI literacy, such as those of Weber et al. (2023), can be helpful for defining universal AI literacy requirements, but we suggest that they require the examination of specific AI aspects

that need to be itemized in a scale dependent on specific AI tasks. Researchers may see AI literacy as a toolbox to identify relevant areas for their specific investigation cases and then look within that context at scales that already exist to combine them for a specific measurement instrument. Operationalizing the toolbox requires fostering interdisciplinary literacy development and perhaps combining measures to bridge the knowledge gaps between different stakeholders. As a consequence, researchers and practitioners aiming to quantify AI literacy should explore empirical applications of the toolbox across various stakeholder groups. We therefore suggest:

> **Recommendation 2**: *Conceptualize AI literacy as a "toolbox" and understand it as a flexible framework that can be adapted to various human-AI interaction contexts in the sense of shared responsibility*.

5.2.2 Addressing the Skilling Problems

Concerning the *skilling problem*, the findings suggest that knowledge workers require both AI literacy and ethical literacy. However, whether knowledge workers possess ethical literacy depends on their ability to reflect critically upon outcomes of human-AI interaction, which depends on their personal dispositions, such as their motivations, moral values, and sense of responsibility.

Some knowledge workers may have certain personal dispositions that help or motivate them to critically reflect on outcomes based on their moral values. As the findings suggest, certain moral values can be fostered, and ethical literacy can be improved to some extent through training, but perhaps not in the same way as training in knowledge or hard skills. Thus, our interviewees advise hiring "the right employees" by searching for those who bring such moral values to the organization. This could mean that hiring decisions should be based more on applicants' moral values, implying ethical literacy, than on hard skills, as also suggested in the literature (e.g., Watson et al., 2021).

Further, there is a need for continuous learning and training (e.g., Chao et al., 2021; Jaiswal et al., 2022). Since AI technologies are continually developing, certain technical skills can become outdated quickly. Thus, knowledge workers need to possess a drive for continual learning. Consequently, the findings suggest that managers need to reconsider their recruiting and personnel skilling strategies when searching for new employees for human-AI interaction, which leads to the following recommendation on how to deal with the skilling problem in organizations:

> **Recommendation 3**: *Recruit for moral values, drive for continuous learning, and focus on organizational training that aims to strengthen these, rather than training only in AI-related knowledge and hard skills*.

While organizations, educators, and policymakers face the challenge of limited resources (Mom et al., 2019), it is crucial to focus on developing the most relevant abilities for specific roles, avoiding deskilling of abilities relevant to responsible human-AI interaction, and allocating resources strategically. Rather than adopting a one-size-fits-all approach, organizations should tailor AI literacy initiatives to the needs of different knowledge workers to maximize impact. Understanding the concept of AI literacy as a toolbox can help in conducting targeted skill development. AI literacy programs should foster interdisciplinary collaboration between different stakeholder groups, enabling them to share knowledge and experiences. Rather than aiming for universal AI literacy, organizations and governments can leverage transactive memory systems that emphasize "who knows what" rather than "everyone knows everything" (Wegner, 1987). Also, organizations can foster role-specific training (M6, H3) and discussions where different stakeholders with diverse abilities are involved in AI development (Asatiani et al., 2020). This may help distribute cognitive labor effectively by structuring knowledge-sharing networks in which individuals specialize in different areas of expertise while remaining aware of the expertise of others to collaborate. Therefore, our recommendation is:

> **Recommendation 4**: *Implement targeted and role-specific AI literacy training by assessing the most critical abilities required for different professional roles and fostering learning from one another*.

5.3 Limitations

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

First, our analysis does not aim to deliver a comprehensive overview of all aspects of AI literacy. Rather, its core purpose is to offer a fresh perspective on how the concept of AI literacy should be rethought, aiming for responsible human-AI interaction. We advocate for moving beyond a purely efficiency-driven approach and instead highlight the need for AI literacy to foster responsible interaction by considering unintended outcomes and inequalities. This broad understanding emphasizes that knowledge workers must be equipped not only with technical know-how but also with the competence to navigate AI systems in ways that are ethically responsible and attuned to the specific contexts in which they operate. For future research, it could be relevant to investigate in more detail and more specific use cases to reveal more practical implications in specific contexts.

Moreover, future research could explore how the role-ability fit varies across different AI application contexts, industries, and organizational settings to examine mechanisms through which organizations can foster the role-ability fit.

Second, the empirical results are restricted to five groups of knowledge workers; however, lawyers and other legal practitioners may also be relevant stakeholder groups when considering literacy for responsible human-AI interaction. Given the initial insights into different use cases in which teaming knowledge workers and AI occurs, future research could enrich the discussion by involving stakeholders from the public domain (government, law) and specific industries. In addition, the progress of AI is widely discussed in the media and academic discourse, which may lead to certain AI narratives that influence individuals' thinking about human-AI interaction. Future longitudinal research could capture the impact of AI developments on what are seen as the skills required for responsible human-AI interaction.

Third, a frequently raised limitation of qualitative research is its restricted generalizability, as the inherently subjective perspectives of interview participants can pose challenges in extending the findings to wider populations (Sarker et al., 2013). To address this, we encourage future research to examine the propositions put forward in our study through quantitative approaches and across diverse domains, use cases, and AI systems. We position our findings as a conceptual contribution aimed at rethinking AI literacy in the context of fostering responsible human-AI interaction.

6 Concluding Comments

The adage "with great power comes great responsibility" applies today as the widespread use of increasingly (semi-) autonomous AI applications at work and in daily life leads to novel human-AI interaction. By analyzing the current literature, we highlight the relevance of rethinking conceptualizations of AI literacy that go beyond enhancing the efficiency and performance of human-AI interactions aimed at fostering abilities to achieve responsible human-AI interaction. By relying on 28 interviews, we aim to increase our understanding of what shapes AI literacy for responsible interactions of knowledge workers with AI.

Our qualitative study contributes to IS research by providing empirical insights into different use cases of human-AI interaction by reflecting on different knowledge workers' perspectives. We illustrate that different use cases require a unique set of AI literacy components to responsibly address potential harms associated with AI while effectively harnessing its benefits. This approach also aids in identifying potential problems that could lead to the exclusion of certain groups, organizations, or nations, thereby helping to reduce the digital divide in human-AI interactions and improving overall welfare, as advocated by Alexopoulou (2024) and Celik (2023). We aim to inform future conceptualizations of AI literacy by offering insights into how the concept may be redefined and the factors that influence its development.

The study's findings highlight that responsible human-AI interaction depends significantly on personal dispositions, which influence whether individuals feel responsible for AI-driven outcomes, potentially creating responsibility gaps. It is recommended to foster a shared sense of responsibility among knowledge workers by fostering ethical dialogue about appropriate AI use across different tasks and contexts. As AI systems become more powerful and complex, a deeper technical understanding is required, while greater usability demands less technical understanding but stronger domain knowledge and communication skills. To address this, AI literacy should be conceptualized as a flexible "toolbox" adaptable to varying interaction contexts, supporting informed and ethical decision-making. Furthermore, responsible interaction with AI calls for understanding multi-literacies, emphasizing that ethical literacy and AI literacy are interdependent and should be strengthened through recruitment for moral values, continuous learning, and targeted organizational training. Finally, to counteract the risks of deskilling from increased AI usability, organizations should implement role-specific AI literacy training based on systematic assessments of critical skill needs to guarantee a role-ability fit.

Finally, we suggest that to speak of *successful* human-AI interaction, we must aim for *responsible* human-AI interaction. By providing future research directions incorporating digital responsibility discourse within AI literacy considerations, and by outlining recommendations for organizational strategies to foster responsible human-AI interaction, we hope this work offers a valuable starting point to reflect upon AI literacy for moving toward *responsible* human-AI interaction in organizations.

Appendix 1: Literature review procedure

We conducted a theoretical literature review to identify what constitutes AI literacy for human-AI interaction according to prior research and explain the initial interdependencies of identified components. A theoretical literature review is appropriate for developing conceptualizations by identifying, describing, and transforming relevant objects into a novel structure. Since theoretical reviews aim at building an explanation through a structured and systematic research process to transform constructs identified in previous work (Paré et al., 2015), the synthesis and reorganization

of AI-related research ensure the development of a holistic view of the ideas of different scholars (Burton-Jones et al., 2021). MacInnis (2011) suggests synthesizing existing theory and transforming it into a novel structure in a conceptualization to "accommodate extant knowledge; explain puzzling or inconsistent findings; [and] reveal novel insights" (p. 138).

Following Templier and Paré (2018), we searched for relevant literature by searching the following terms in titles, abstracts, and keywords: "AI literacy," "AI capabilities," "AI abilities," "AI competencies," "AI knowledge," and "AI skills." We searched these databases: ACM Digital Library (ACM), AIS Electronic Library (AIS), Business Source Ultimatum via EBSCOhost (BSU), IEEE Explore (IEEE), JSTOR, Science Direct (SciDir), and Web of Science (WoS). Due to the topic's novelty, we wanted to ensure that we could consider the latest relevant conference articles, so we did not restrict our search to peer-reviewed literature or specific quality criteria (e.g., rankings); we did, however, limit our search to articles published between 2017 and 2023.

Our final search, conducted in June 2023, yielded 2,683 overall hits across all databases without the use of database filter options, 1,868 hits across all databases with the use of the database filter option that excluded non-English articles automatically ($n=22$), and articles published before 2017 ($n=793$), and 1,666 unique hits after the elimination of duplicates ($n=202$). In addition to excluding non-English articles automatically, we applied further format criteria as detailed in Table 3.

We applied further relevance criteria. To increase the quality of relevance screening, we randomly chose a small subsample of the articles ($n=37$) for parallel training purposes of the coders before conducting relevance screening. The relevance criteria were revised through subsequent discussion. The 763 articles that did not fit our relevance criteria were excluded, and 37 articles remained in the final sample. Table 4 provides an overview of the relevance screening:

We also conducted backward and forward searches and checked all articles against the format and relevance criteria described above. The remaining articles were subjected to further rounds of backward and forward searches. We repeated the process until no further relevant articles could be identified. In total, we had three rounds of backward and three rounds of forward search, resulting in 118 hits. We added nine articles resulting in 46 articles as our final sample. Figure 4 is an overview of the search process based on Templier and Paré (2018).

For the systematic synthesis of our final sample, we followed a fourth-step structure as suggested by MacKenzie to avoid "poor construct conceptualization" (2003, p. 324), which can lead to misspecifications during theory development. The first step is about defining constructs of interest and the identification of similarities among groups of capabilities, competencies, and skills (representing AI literacy aspects). The second step is about synthesizing similar expressions of AI literacy aspects and their corresponding building blocks. The third concerns thinking about

Table 3 Format screening

Database	Before 2017	Non-English	Non-scientific	Incomplete	Total
Definition	*Articles being published before 1st January 2017*	*Articles that are not completely written in English*	*Grey literature, book chapters, book reviews, dissertations, tutorials, bachelor's or master's theses, workshops, posters, letters of editors, special issues, symposiums, talks, and magazines*	*Short papers or research-in-progress papers, and extended abstracts*	-
ACM	55	0	26	17	98
AIS	30	0	25	13	68
EBSCO	159	22 (filter)	520	0	701
IEEE	266	3 (manual)	24	1	294
JSTOR	5	0	163	0	168
SciDir	55	0	11	0	66
WoS	223	2 (manual)	51	10	286
Others	-	-	-	-	-
Total	**793** (with database filter)	**22** (with database filters) **5** (manual screening)	**820** (manual screening)	**41** (manual screening)	**1,681** (total) **866** (manual screening) **815** (with database filters)

Table 4 Relevance screening

Database	No AI focus	No AI literacy focus	No human-AI interaction focus	Total
Definition	*Technologies mentioned in the articles must be characterized by a higher level of complexity and interactivity compared to previous technologies based on AI's self-learning abilities and increased autonomy, as illustrated in the introduction*	*Articles must provide dedicated contributions regarding the components of AI literacy (aspects, building blocks, or interrelations with other literacies)*	*Articles must consider both parties (human and AI teammates) being involved in human-AI interaction. Hence, articles solely focussing on the design of AI by addressing AI abilities or solely focussing on managing or enhancing human abilities without considering the intersection with AI are not included*	-
ACM	1	34	12	47
AIS	1	77	41	119
EBSCO	35	116	122	273
IEEE	0	33	27	60
JSTOR	10	23	6	39
SciDir	1	30	22	53
WoS	11	89	72	172
Others	-	-	-	-
Total	**59**	**402**	**302**	**763**

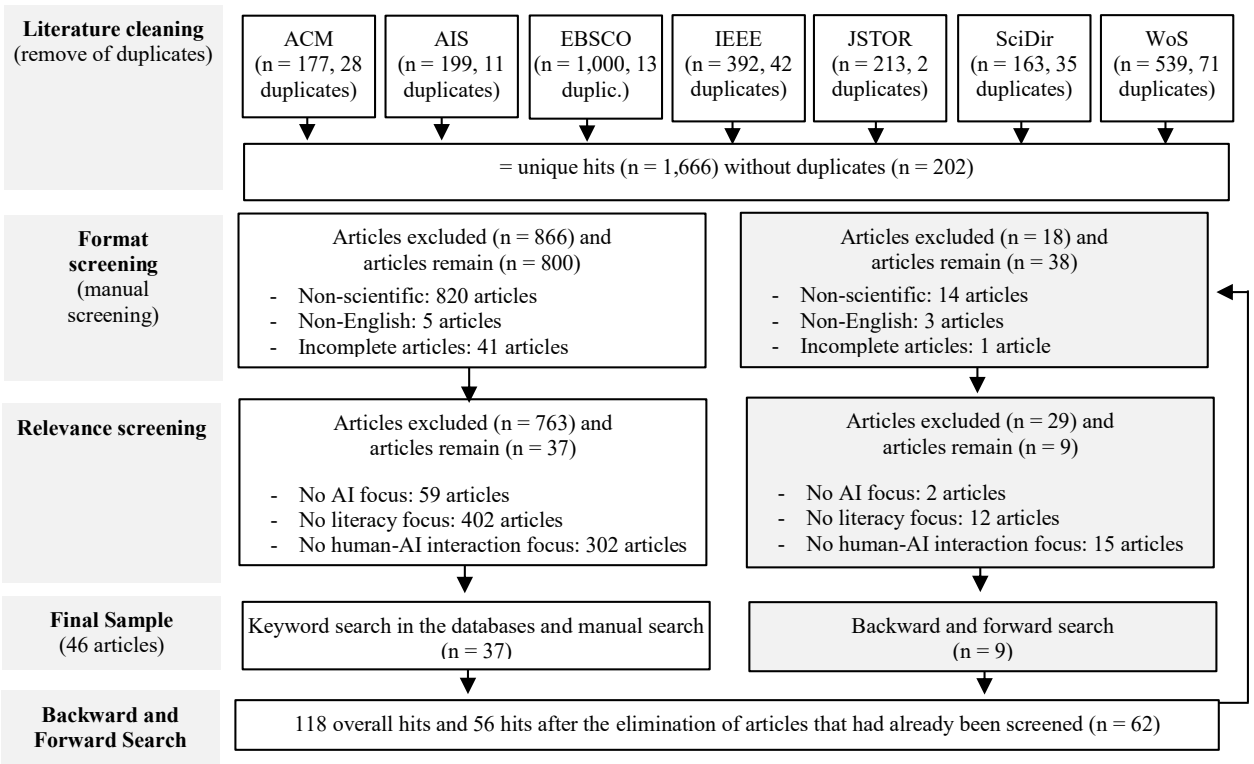


Fig. 4 Literature search process

adequate measures representing the constructs and the relation between measures and constructs. The fourth step is about the identification of relationships among constructs and the classification of those. We aimed to contribute to the design validity through our systematic and transparent description of the steps of our theoretical literature review. We aimed to increase measurement and analytical validity by having several discussion rounds amongst at least two researchers when synthesizing constructs until consensus was reached and by reporting interrater reliability using Cronbach's Alpha (McHugh, 2012). We mainly followed deductive reasoning when synthesizing constructs based on the theoretical background to increase construct and inferential validity.

Table 5 Final sample

Nbr	Autor(s)	Year	Title	Outlet	Source
1	Alekseva et al	2021	AI Adoption and Firm Performance: Management versus IT	Academy of Management	Keyword Search
2	Ali et al	2021	Children as Creators, Thinkers and Citizens in an AI-Driven Future	Computers and Education: Artificial Intelligence	Keyword Search
3	Alsheiabni et al	2019	Towards an Artificial Intelligence Maturity Model: From Science Fiction to Business Facts	Pacific Asia Conference on Information Systems (PACIS)	Keyword Search
4	Anton et al	2020	The Humans behind Artificial Intelligence – an Operationalisation of AI Competencies	European Conference on Information Systems (ECIS)	Keyword Search
5	Cetin-damar et al	2021	Explicating AI Literacy of Employees at Digital Workplaces	IEEE Transaction on Engineering Management	Keyword Search
6	Chai et al	2020	An Extended Theory of Planned Behavior for the Modelling of Chinese Secondary School Students' Intention to Learn Artificial Intelligence	Mathematics	Keyword Search
7	Carolus et al., 2023	2023	Digital interaction literacy model – Conceptualizing competencies for literate interactions with voice-based AI systems	Computers and Education: Artificial Intelligence	Keyword Search
8	Chao et al	2021	Knowledge of and Competence in Artificial Intelligence: Perspectives of Vietnamese Digital-Native Students	IEEE Access	Forward Search 1 st round
9	Casal-Otero et al	2023	AI literacy in K-12: a systematic literature review	International Journal of System Education	Keyword Search
10	Dai et al	2020	Promoting Students' Well-Being by Developing Their Readiness for the Artificial Intelligence Age	Sustainability	Keyword Search
11	Dai et al	2023	Collaborative construction of artificial intelligence curriculum in primary schools	Journal of Engineering Education	Keyword Search
12	George	2021	Redesign of an Elective Introductory Artificial Intelligence Course in a Credit-Limited Computer Science Curriculum	Journal of Computing Sciences in Colleges	Keyword Search
13	Hemmet et al	2023	AI in the Public Eye: Investigating Public AI Literacy Through AI Art	ACM Conference on Fairness, Accountability, and Transparency	Keyword Search
14	Hermann	2022	Artificial Intelligence and Mass Personalization of Communication Content— An Ethical and Literacy Perspective	New Media & Society	Keyword Search
15	Hou et al	2021	Study on the Competency Model Construction for Industrial Designers under Artificial Intelligence Technology	International Conference on Electronic Business (ICEB)	Keyword Search
16	Jaiswal et al	2021	Rebooting Employees: Upskilling for Artificial Intelligence in Multinational Corporations	The International Journal of Human Resource Management	Backward Search 2nd round
17	Kaspersen et al	2021	VotestratesML: A High School Learning Tool for Exploring Machine Learning and its Societal Implications	FabLearn Europe/MakeEd	Keyword Search
18	Kim et al	2021	Why and What to Teach: AI Curriculum for Elementary School	Conference on Innovative Applications of Artificial Intelligence	Keyword Search
19	Kong et al	2023	Evaluating an Artificial Intelligence Literacy Programme for Developing University Students? Conceptual Understanding, Literacy, Empowerment and Ethical Awareness	Educational Technology & Society	Keyword Search
20	Laupichler et al	2022	Artificial Intelligence Literacy in Higher and Adult Education: A Scoping Literature Review	Computers and Education: Artificial Intelligence	Forward Search 1 st round
21	Laupichler et al	2023	Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy	Computers and Education: Artificial Intelligence	Keyword Search
22	Leyer & Schneider	2021	Decision Augmentation and Automation with Artificial Intelligence: Threat or Opportunity for Managers?	Business Horizons	Keyword Search
23	Long & Magerko	2020	What is AI Literacy? Competencies and Design Considerations	Conference on Human Factors in Computing Systems (CHI)	Keyword Search
24	Long et al	2021	Co-Designing AI Literacy Exhibits for Informal Learning Spaces	ACM Human–Computer Interaction	Keyword Search
25	Long et al	2022	Family Learning Talk in AI Literacy Learning Activities	Conference on Human Factors in Computing Systems (CHI)	Keyword Search

Table 5 (continued)

Nbr	Autor(s)	Year	Title	Outlet	Source
26	Luo et al	2023	Aladdin's Genie or Pandora's Box for Early Childhood Education? Experts Chat on the Roles, Challenges, and Developments of ChatGPT	Early Education and Development	Keyword Search
27	Markauskaiet et al	2022	Rethinking the Entwinement Between Artificial Intelligence and Human Learning: What Capabilities do Learners Need for a World with AI?	Computers and Education: Artificial Intelligence	Forward Search 1 st round
28	Mikalef & Gupta	2021	Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on its Impact on Organizational Creativity and Firm Performance	Information & Management (IM)	Keyword Search
29	Mikalef et al	2021	Artificial Intelligence as an Enabler of B2B Marketing: A Dynamic Capabilities Micro-Foundations Approach	Industrial Marketing Management	Keyword Search
30	Ng et al	2021	Conceptualizing AI literacy: An Exploratory Review	Computers and Education: Artificial Intelligence	Keyword Search
31	Ng et al	2022	Using Digital Story Writing as a Pedagogy to Develop AI Literacy Among Primary Students	Computers and Education: Artificial Intelligence	Keyword Search
32	Ng et al	2023	Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world	Educational Technology Research and Development	Keyword Search
33	Pinski & Benlian	2023	AI Literacy—Towards Measuring Human Competency in Artificial Intelligence	Hawaii International Conference on System Sciences (HICSS)	Forward Search 1 st round
34	Pinski et al	2023	AI Knowledge: Improving AI Delegation through Human Enablement	CHI Conference on Human Factors in Computing Systems	Keyword Search
35	Sanusi et al	2022	Investigating Learners' Competencies for Artificial Intelligence Education in an African K-12 Setting	Computers and Education Open	Backward Search 2nd round
36	Sanusi et al	2022	The Role of Learners' Competencies in Artificial Intelligence Education	Computers and Education: Artificial Intelligence	Forward Search 1 st round
37	Su & Yang	2023	Artificial Intelligence (AI) literacy in early childhood education: an intervention study in Hong Kong	Interactive Learning Environments	Keyword Search
38	Su & Zhong	2022	Artificial Intelligence (AI) in Early Childhood Education: Curriculum Design and Future Directions	Computers and Education: Artificial Intelligence	Keyword Search
39	Su et al	2022	A Meta-Review of Literature on Educational Approaches for Teaching AI at the K-12 Levels in the Asia–Pacific Region	Computers and Education: Artificial Intelligence	Keyword Search
40	Van Brummelen et al	2021	Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools	Conference on Innovative Applications of Artificial Intelligence	Keyword Search
41	Verma et al	2022	An Investigation of Skill Requirements in Artificial Intelligence and Machine Learning Job Advertisements	Industry and Higher Education	Forward Search 1 st round
42	Wang et al	2022	Measuring User Competence in Using Artificial Intelligence: Validity and Reliability of Artificial Intelligence Literacy Scale	Behavior & Information Technology	Keyword Search
43	Watson et al	2021	Will AI Ever sit at the C-Suite Table? The Future of Senior Leadership	Business Horizons	Keyword Search
44	Yang, W	2022	Artificial Intelligence Education for Young Children: Why, What, and How in Curriculum Design and Implementation	Computers and Education: Artificial Intelligence	Keyword Search
45	Yang, S	2022	A Systematic Literature Review on the Disruptions of Artificial Intelligence within the Business World: in Terms of the Evolution of Competences	Conférence de l'AIM	Forward Search 1 st round
46	Zhao et al	2022	Developing AI Literacy for Primary and Middle School Teachers in China: Based on a Structural Equation Modeling Analysis	Sustainability	Keyword Search

Appendix 2: Overview of participants per stakeholder group

Stakeholder	Ref. ID	Position	Organization	Expertise	#employees
Recruiters & HR people ($n=5$)	H1	Recruiter	Technology Provider	Preselection of candidates with AI	>135.000
	H2	Head of IT Transformation	Automotive Supplier	AI recruiting agent; autonomous manufacturing	>30.000
	H3	Recruiter	Insurance Company	Candidate assessment with AI; Employer branding with generative AI	>30.000
	H4	HR Consultant	Head-hunter Agency	Candidate assessment with AI	<10
	H5	CEO	Hiring Agency	AI recruiting agent	<50
Managers ($n=6$)	M1	CEO	Ad & Media Agency	Content creation with ChatGPT and Dall-E	<50
	M2	Head of People & Culture	Insurance Company	Damage assessment with AI; AI recruiting agent	>10.000
	M3	Marketing Manager	Renewable Energy Provider	AI drone operations; content creation with ChatGPT	<500
	M4	CEO	Industrial Automation	AI detection in manufacturing for predictive maintenance	<100
	M5	Board of Directors Assistance	Pharmaceutics	Data analytics for performance evaluation; digital skiing teacher	>1000
	M6	Marketing Manager	News & Service Platform	Content creation in journalism, predictions in marketing	>2000
Developers & Data Scientists ($n=6$)	D1	Software Engineer	AI Software start-up	GitHub and Microsoft Copilot for programming	>100
	D2	Software Developer	Software Company	AI-based decision support in finance	>220.000
	D3	Data Scientist	Car Manufacturer	AI simulation and image generation; ChatGPT for information search	>150.000
	D4	Lead Developer	Technology Provider	Smart home to optimize processes	>300.000
	D5	Software Developer	Tax and Audit Consulting	Algorithms for predictions; GitHub and Microsoft Copilot for programming	>300.000
	D6	IT Security Developer	Manufacturing comany	AI bots used for cybersecurity and threat mitigation	>80.000
Educators & Researchers ($n=5$)	R1	AI analyst & Researcher	AI Research Lab	Analyse social robots in care work; robots in manufacturing	>10
	R2	Researcher	Business Analytics Institute	Educate students on explainable AI and governance mechanisms	>50
	R3	AI research & Lecturer	University	Anomaly detection & intelligent manufacturing	>10
	R4	Project Lead	Institute for Vocational Training	Learning records platform that supports personalized studying	>20
	R5	Lecturer & Researcher	University	Mood detection through face analysis	>10
AI providers & sales personnel ($n=6$)	S1	Product Owner	Medical Technology Services	Smart Manufacturing & Anomaly Detection through data analytics	>300.000
	S2	Sales Lead	Technology Provider	Programming with GitHub Copilot	>300.000
	S3	Practice Director	AI Software Start-up	anomaly detection in manufactoring	>100
	S4	Innovation Architect	Venture Capital	Information search with ChatGPT; matching tool in real estate market	>10.000
	S5	Enterprise Account Executive	Cloud Services	Fraud detection in ticket sales	>500.000
	S6	Co-Founder	AI Startup	App that does facial rcognition of animals	>5

Appendix 3: Semi-structured interview guide

Section	Tasks and its underlying questions
Entry	• Welcome the interviewee • Ask the interviewee to test the microphone and to switch on the camera (voluntary) • Shortly introduce the researchers and its functions • Shortly repeat the project background and aim ○ Note: The abbreviation AI stands for Artificial Intelligence. (To avoid influencing, we deliberately do not use a definition as a basis, we ask for the self-understanding in the second section of our interview.) • Explain the structure of the interview • Explain what happens with the data after the interview and explain that solely the audio will be recorded and deleted after being transcribed • Ask if there are questions regarding the process or the interview beforehand • Ask if they agree to be recorded and start the recording
i) Introduction	We start with the first block of our interview, where we would like to know some information about the interviewee's background and functions: • In what type of company or institution are you currently employed? [Context: business model, industry, and corporate culture] ○ What is the business model and vision behind it? ○ How would you describe the culture in your work environment, esp. with regard to technology use? [Attention—check: to which unit/level does the answer refer: UN vs. team] ○ What is the role of AI in the company? • What is your current position, and how long have you held it? [Assign job history & educational background] ○ What work experience did you have before your current position? • What are your current primary responsibilities? [Context: Activities] ○ Do you have responsibility for individuals? If so, for how many and to what extent? ○ Do you have responsibility for technologies (e.g., maintenance, implementation decision)? If yes, what types of technologies and to what extent?
ii) AI definition and use case(s)	• What is your understanding of AI? Please briefly describe what impulses, examples, or characteristics come to mind when you hear the buzzword AI ○ How would you differentiate between Human Intelligence and AI? • Please describe at least one example where AI-based technologies are used for private tasks in daily life or working situations in your daily business. Please clearly describe the task, its purpose, the type/group of individuals that conduct this task in cooperation with AI-based technologies, and the situation in which the task is executed [Explanation: These can be AI-based systems with which you interact directly (in the role of the person using them) or with which others (e.g., employees, students, customers) interact and with which you are therefore more indirectly involved (e.g., responsibility, implementation, training). These can be AI-based technologies controlled by the organization (i.e., initiated by the subordinate), as well as the uncontrolled and selective use of such technologies.] ○ Specific task of human-AI interaction ○ Purpose/Aim (possible outcome variable(s)) of human-AI interaction ○ Type of Individual/Group of People conducting the task ○ Type of the AI or the technology in which the AI is embedded ○ Situation (e.g., frequency, routinized or non-routinized task)

Section	Tasks and its underlying questions
iii) Context-specific requirements	Questions per stakeholder group: • Recruiters: ○ What can HR do to improve human-AI interaction? ○ What abilities are organizations searching for when hiring individuals to work on AI projects? ○ What abilities do recruiters need for future candidate acquisition and selection with AI? • Managers: Please describe the ideal workforce: ○ What abilities are needed from the workforce in this/those example/s to deal with AI (even though this does not represent reality)? ○ What are the minimum abilities individuals need to accomplish the underlying task in your mentioned example/s? ○ If you would need to hire someone for your specific task mentioned, for which abilities would you ask? • Developers/Data Scientists: ○ As an AI developer, what can you do to improve the human-AI interaction of later users? ○ What are the abilities developers need when designing AI or training AI? ○ What abilities do you expect from users? • Educators/Researchers: ○ What abilities does a society or state require for human-AI interaction? ○ How would you describe the future abilities of AI-literate citizens? • Sales/Product Managers: ○ What abilities do organizations require when aiming to deploy AI? ○ What are the minimum abilities organizations must have to implement AI-based technologies successfully? ○ If you would need to hire a team member in your sales team for your specific task mentioned, for which abilities would you ask? ○ How would you describe the ideal client organization for successful and responsible AI project implementation? How is their leadership team?
v) Context-specific barriers	[Explanation] The interviewee mentioned several abilities for specific situations of human-AI interaction. Now, they should describe what has been done, what is done, and what should be done (from individuals, the organization, or society) to establish or foster those abilities Questions per stakeholder group: • Recruiters: ○ What do you think hinders an individual's development of abilities needed for interaction between humans and AI? ○ What can the single employee do to foster capabilities in human-AI interaction? • Managers: ○ What do you think hinders a workforce's development of abilities needed for interaction between humans and AI? ○ What do you think hinders an organization's development of abilities needed for AI deployment? ○ What can organizations do to foster the development of relevant abilities required for deploying and using AI in organizations? • Developers/Data Scientists: ○ What do you think is hindering the development of abilities of individuals that are needed for AI development? ○ What can the developers do to foster capabilities in human-AI interaction? • Educators/Researchers: ○ What do you think hinders a society's development of abilities needed for interaction between humans and AI? ○ What can society or a state do to ensure that humans and AI-based technologies work together? • Sales/Product Managers: ○ What do you think hinders an organization's development of abilities needed for AI deployment? ○ What can organizations do to foster the development of relevant abilities required for human-AI interaction and prepare an organization for human-AI interaction?
Closing	• Stop the recording and clarify the end of the interview • Explain what will happen in the next step • Ask if there are any questions • Ask how they find the interview (feedback for researchers) • Ask for suggestions on what else should be considered in the current study, to help discover any potential incompleteness • Thank the interviewees

Appendix 4: Snapshot of coding

Part of the Conceptualization	Code	Coding Rule/Definition	Coding Example	Descriptive Statistics
Literacies	Ethical literacy	Articulate ethical viewpoints and reflect moral values	"we had discussions with the software providers on how to make sure that the AI assessment is objective. For example, the AI might reject an applicant due to a foreign name. The question is, how do we notice that? Therefore, it is important that you have this objective view somehow (dual control) and that you can take this step back to analyse what's happening there. Usually, you have it mapped through hierarchies or other decision-making policies. You can't have one person deciding alone. You have at least two humans deciding together. And maybe this can be transferred to the AI by having a technical solution for dual control within decision-making." (M2)	45.16% of 62 overall codes
	Writing literacy	Compose and read text for communication and learning	"How do I communicate a problem? [That's] important both in communication among developers and with the customer or with AI because if I can't say cleverly what my problem is, then no AI tool in the world will help me." (D1)	3.23% of 62 overall codes
	Legal literacy	Awareness of legal issues based on knowledge about law and human rights	"People can be checked at work—with or without AI. How many parcels they have delivered? How long was someone in the bathroom, and how often did someone make a break? Employee A delivered 100, and Employee B delivered 300 packages. Is it ethical or unethical to compare that? This legal situation is the same with or without AI." (M5)	8.06% of 62 overall codes
	Data literacy	Ability to collect, analyze, structure, and interpret data	"That means many people assume building the model itself is extremely demanding and requires a lot of expertise. But in fact, it's the other way around: 90% of the work lies in preparing and identifying the data, which is the decisive part." (S3)	27.42% of 62 overall codes
	Digital literacy	Ability to interact effectively with digital tools such as information and communication technologies	"The necessary prerequisite is being able to operate digital technologies, that is, having basic IT skills. You can use a smartphone, a laptop, set up an internet connection, and similar things" (R2)	6.45% of 62 overall codes
	Algorithmic literacy	understanding and appropriately utilizing algorithms	"specifically with Google—we know they're skimming something off, but we have no idea at which point something happens and what exactly is being collected. Similar to Outbrain, we don't know that at all. For me, the first step would be to understand what is going where, and at what stage anything is being transferred at all." (M6)	9.68% of 62 overall codes

The 'Ref. IDs' such as "M2" or "D1" correspond to the participants listed in Appendix 2

Appendix 5: Findings regarding building blocks and aspects

Use Cases	Ref. ID	Building blocks	Aspects
Decision-making tasks • Recruiting and pre-selection of candidates (H1, H5, M2, M6) • Candidate assessment in recruiting (H2, H3, H4)	• H1 • H2 • H3 • H4 • H5	Personal dispositions	• open mindset (M2, H1, H4) • moral intuition (M2) • sense of responsibility (H5, D2) • ethical awareness (D2, R2, M6) • self-efficacy (R2, H4)
• Decision-support in financial consulting (D2) • Explainable AI (R2)	• M2 • D2 • R2	Knowledge	• know AI's limits (H1, H2) • understanding AI's implications (D2) • understand AI's functioning (M2, R2, D2, R2) • understand AI's underlying mechanisms (H1, H2, R2) • domain specific knowledge (H2)
		Skills	• collaboration (M2) • communication (M2) • critical reflection (H1, H2, H3, D2, R2)
Creation and information tasks • Content creation (H3, D1, D3, M1, M3, M6) • Information search (D3, S4) • Programming with generative AI (D1, D5, S2)	• H3 • M1 • M3 • M6 • D1 • D3	Personal disposition	• patience (S2, D1) • open mindset (S2) • healthy skepticism (D5) • moral intuition (M6) • tolerance for failing (S2) • cultural sensitivity (M3)
	• D5 • S2 • S4	knowledge	• know AI's limits (D1) • understanding AI's underlying mechanism (D3, S2) • understanding AI implications (M3) • domain specific knowledge (M6)
		skills	• communication skills (S2, D1, D3, H3, M6) • aligning AI with business proccesse (D3, M3)
Prediction tasks • Data analytics and algorithmic prediction (D1, D5, M6)	• M4 • M6 • D1 • D5	Personal disposition	• self-learning abilities (M4, D5, S5, S6) • ethical understanding and moral values (M6, D1, D5, S5, M4, R5) • sense of responsibility (M6, S6)
• Fraud detection (S5) • Predictive maintenance (M4) • Mood detection and facial recognition (S6, R5)	• R5 • S5 • S6	knowledge	• legal understanding (R5) • understand AI's underlying mechanism (M4, S5, D1) • domain specific knowledge (S6)
		Skills	• critical reflection (D1, D5, S6, M4, M6, R5) • communication skills (M4, M6, D1) • aligning AI with business proccesse (D5, M6, S6)
Optimization tasks • Data analytics and performance evaluations (M5) • Smart home (D4) • AI drone operations (M3) • Anomaly detection and damage assessment (M2, R3, S1, S3) • Bots for threat mitigation (D6)	• M2 • M3 • M5 • D4 • D6 • R3 • S1 • S3	Personal disposition	• openness (M2, D4, S1) • innovative spirit (M2, M3, D4, S1, S3) • patience (M2, D4) • ethical awareness (M2, R3, S3) • sense of responsibility (M3, D4, R3) • cultural sensibility (M5, R3, S1) • collaboration (M5, D4, S3) • self-learning ability (D4, D6, S1, S3) • acumen for business decisions (R3, S1)
		Knowledge	• knowing AI governance (R3) • tool expertise (M3, M5, S3) • statistical basics (M2, M5, S3) • understanding AI's implications (M3, M5, D4, D6, S3) • understand AI's basic functioning (M3, M5, D6, R3, S1, S3)

Use Cases	Ref. ID	Building blocks	Aspects
		Skills	• programming skills (D4, R3, S3) • communication skills (D6, R3, S3) • aligning AI with business processes (S3) • critical reflection (M5, M2, D4, R3, S3)
Physical tasks • Healthcare robots in elderly care (R1)	• R1	personal disposition	• cognitive flexibility (R1) • ethical awareness (R1)
• Robots in manufacturing for material handling (R1)		knowledge	• understand AI's functioning (R1)
		skills	• programming skills (R1) • critically evaluate AI's impact (R1)
Communicative and supportive tasks	• R4 • S4	knowledge	• know AI's functioning (S4, R4) • understanding AI's implications (R4)
• Chatbots for customer support • Teaching agents (R4, M5) • Matching algorithms for customer preferences (S4)		skills	• tool expertise (S4) • communication skills (S4) • aligning AI with business processes (R4)

The 'Ref. IDs' correspond to the participants listed in Appendix 2

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

Authors' Contributions Conceptualization: N.P., T.H., O.P.; methodology: N.P., T.H.; research: N.P., T.H.; conducting interviews: N.P., T.H.; data analysis: N.P., T.H.; writing – original draft preparation: N.P., T.H.; writing – review and editing: N.P., T.H., O.P.; supervision: O.P.; project administration: N.P.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate All interview participants took part in the study voluntarily, and all ethical standards were adhered to during the conduct and analysis of the interviews.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107–136. https://doi.org/10.2307/3250961

Alexopoulou, S. (2024). Paradigm shift: Exploring the impact of digital technologies on the welfare state through a systematic literature review. *Social Policy and Administration*. https://doi.org/10.1111/spol.13054

Alexopoulou, S., Åström, J., & Karlsson, M. (2022). The grey digital divide and welfare state regimes: A comparative study of European countries. *Information Technology & People, 35*(8), 273–291. https://doi.org/10.1108/ITP-11-2020-0803

Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence, 2*, Article 100040. https://doi.org/10.1016/j.caeai.2021.100040

Alsheiabni, S., Cheung, Y., & Messom, C. (2019). Towards an artificial intelligence maturity model: from science fiction to business facts. *Pacific Asia conference on information systems (PACIS)*.

Anton, E., Behne, A., & Teuteberg, F. (2020). The Humans Behind Artificial In[TEL]lligence – An Operationalization Of AI Competencies. In *European Conference on Information Systems (ECIS)*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

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, 19*(4), 259–278.

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior, 45*(2), 159–182. https://doi.org/10.1002/job.2735

Bawden, D. (2008). Origins and concepts of digital literacy. *Digital Literacies: Concepts, Policies and Practices, 30*, 17–32.

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), 10.

Benlemlih, M., & Bitar, M. (2018). Corporate social responsibility and investment efficiency. *Journal of Business Ethics, 148*, 647–671.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*. https://doi.org/10.25300/MISQ/2021/16274

Boichak, O., Canzonetta, J., Sitaula, N., McKernan, B., Taylor, S., Rossini, P., ..., Stromer-Galley, J. (2019). Beyond the Medium: Rethinking Information Literacy through Crowdsourced Analysis. In T. X. Bui (Ed.), *Proceedings of the 52nd annual hawaii international conference on system sciences*. University of Hawaii at Manoa Hamilton Library ScholarSpace.

Bradley, J. H., Paul, R., & Seeman, E. (2006). Analyzing the structure of expert knowledge. *Information and ManagemeNt, 43*(1), 77–91. https://doi.org/10.1016/j.im.2004.11.009

Burton-Jones, A., Butler, B. S., Scott, S., & Xu, S. X. (2021). Next-generation information systems theorizing: A call to action. *MIS Quarterly, 45*(1), 301–314.

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology, 13*, 836650.

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*, Article 100114.

Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The sociomateriality of information systems. *MIS Quarterly, 38*(3), 809–830.

Celik, I. (2023). Exploring the determinants of artificial intelligence (AI) literacy: Digital divide, computational thinking, cognitive absorption. *Telematics and Informatics, 83*, Article 102026.

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. https://doi.org/10.1109/TEM.2021.3138503

Chao, P.-J., Hsu, T.-H., Liu, T.-P., & Cheng, Y.-H. (2021). Knowledge of and competence in artificial intelligence: Perspectives of Vietnamese digital-native students. *Ieee Access, 9*, 75751–75760. https://doi.org/10.1109/ACCESS.2021.3081749

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*. https://doi.org/10.1007/s10676-021-09616-9

Crewswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE.

Dalkir, K. (2013). *Knowledge management in theory and practice*. Routledge.

DiSessa, A. A. (2018). Computational literacy and \textquotedblthe big picture\textquotedbl concerninc computers in mathematics education. *Mathematical Thinking and Learning, 20*(1), 1–43.

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

Eshleman, A. (2014). *Moral responsibility*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/archives/fall2019/entries/moral-responsibility/

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science, 331*(6018), 721–725. https://doi.org/10.1126/science.1201765

Ferguson, L. W. (1960). Ability, interest, and aptitude. *Journal of Applied Psychology, 44*(2), 126–131.

Ferguson, M. J., Hassin, R., & Bargh, J. A. (2008). Implicit motivation: Past, present, and future. *Handbook of motivation science*, 150–166.

Frezza, S., Daniels, M., Pears, A., Cajander, Å., Kann, V., Kapoor, A., McDermott, R., Peters, A.-K., Sabin, M., & Wallace, C. (2018). Modelling competencies for computing education beyond 2020: A research based approach to defining competencies in the computing disciplines.In *Proceedings Companion of the 23rd Annual ACM conference on innovation and technology in computer science education*, 148–174. https://doi.org/10.1145/3293881.3295782

Gabric, D., & McFadden, K. L. (2001). Student and employer perceptions of desirable entry–level operations management skills. *American Journal of Business, 16*(1), 5–60. https://doi.org/10.1108/19355181200100005

Garrison, D. R., Cleveland-Innes, M., Koole, M., & Kappelman, J. (2006). Revisiting methodological issues in transcript analysis: Negotiated coding and reliability. *The Internet and Higher Education, 9*(1), 1–8. https://doi.org/10.1016/j.iheduc.2005.11.001

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

Gilster, P. (1997). *Digital literacy*. John Wiley & Sons.

Hamilton, V. L., & Hagiwara, S. (1992). Roles, responsibility, and accounts across cultures. *International Journal of Psychology, 27*(2), 157–179. https://doi.org/10.1080/00207599208246873

Hemment, D., Currie, M., Bennett, S. J., Elwes, J., Ridler, A., Sinders, C., Vidmar, M., Hill, R., & Warner, H. (2023). AI in the Public Eye: Investigating Public AI Literacy Through AI Art. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 931–942.

Hermann, E. (2022). Artificial intelligence and mass personalization of communication content—An ethical and literacy perspective. *New Media & Society, 24*(5), 1258–1277. https://doi.org/10.1177/14614448211022702

Heyder, T., & Posegga, O. (2021). Extending the foundations of AI literacy. In *International conference on information systems (ICIS)*.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems, 32*(3), Article 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.

Horneber, D., & Laumer, S. (2023). Algorithmic accountability. *Business & Information Systems Engineering, 65*(6), 723–730.

Jaiswal, A., Arun, C. J., & Varma, A. (2022). Rebooting employees: Upskilling for artificial intelligence in multinational corporations. *The International Journal of Human Resource Management, 33*(6), 1179–1208. https://doi.org/10.1080/09585192.2021.1891114

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial Intelligence and Computer Science in Education: From Kindergarten to University. In *IEEE Frontiers in education conference (FIE)*. 1-9. https://doi.org/10.1109/FIE.2016.7757570

Kathala, K. C. R., & Palakurthi, S. (2024). AI Literacy Framework and Strategies for Implementation in Developing Nations. *Proceedings of the 2024 16th international conference on education technology and computers*, 418–422. https://doi.org/10.1145/3702163.3702449

Kegel, M., & Ghanem, L. (2024). You did that on purpose! An Investigation of the knobe effect in human-robot interactions. In *Proceedings of the 57th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2024.088

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

Kunz, W. H., & Wirtz, J. (2024). Corporate digital responsibility (CDR) in the age of AI: Implications for interactive marketing. *Journal of Research in Interactive Marketing, 18*(1), 31–37.

Lankshear, C., & Lawler, M. (1987). *Literacy, schooling and revolution*. Falmer.

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Laupichler, M. C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence, 4*, Article 100126.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science, 33*(1), 126–148.

Leonardi, P. M. (2012). *Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they related? Do we need them?* (SSRN Scholarly Paper 2129878). https://doi.org/10.2139/ssrn.2129878

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255.

Lin, C. H., Wu, L. Y., Wang, W. C., Wu, P. L., & Cheng, S. Y. (2020). Development and validation of an instrument for AI-Literacy. *3rd Eurasian conference on educational innovation (ECEI)*.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *CHI conference on human factors in computing systems* (pp. 1–16). https://doi.org/10.1145/3313831.3376727

Long, D., Teachey, A., & Magerko, B. (2022). Family Learning Talk in AI Literacy Learning Activities. *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–20.

Luke, C. (1997). *Technological Literacy: Research into Practice Series No. 4*. National Languages and Literacy Institute of Australia.

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

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science, 31*(3), 323–326.

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, 43*(2), 334–354.

Manca, S., Bocconi, S., & Gleason, B. (2021). "Think globally, act locally": A glocal approach to the development of social media literacy. *Computers & Education, 160*, 104025. https://doi.org/10.1016/j.compedu.2020.104025

Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., de Laat, M., Buckingham Shum, S., Gašević, D., & Siemens, G. (2022). Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI? *Computers and Education: Artificial Intelligence, 3*, Article 100056. https://doi.org/10.1016/j.caeai.2022.100056

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-018-3921-3

Martini, F., & Sénéchal, M. (2012). Learning literacy skills at home: Parent teaching, expectations, and child interest. *Canadian Journal of Behavioural Science = Revue Canadienne Des Sciences Du Comportement, 44*(3), 210–221.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended consequences of introducing AI systems for decision making. *MIS Quarterly Executive, 19*(4), 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). Lessons in biostatistics: Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 359–372. https://doi.org/10.1201/b16720-37

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

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information and ManagemeNt, 58*(3), Article 103434. https://doi.org/10.1016/j.im.2021.103434

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

Morandini, S., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., & Pietrantoni, L. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. *Informing Science: The International Journal of an Emerging Transdiscipline, 26*, 39–68.

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., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2021.100041

Ng, D. T. K., Luo, W., Chan, H. M. Y., & Chu, S. K. W. (2022). Using digital story writing as a pedagogy to develop AI literacy among primary students. *Computers and Education: Artificial Intelligence, 3*, Article 100054. https://doi.org/10.1016/j.caeai.2022.100054

Nordbye, G. H. H., & Teigen, K. H. (2014). Being responsible versus acting responsibly: Effects of agency and risk taking on responsibility judgments. *Scandinavian Journal of Psychology, 55*(2), 102–114. https://doi.org/10.1111/sjop.12111

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*. https://doi.org/10.1177/14614448231182662

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 and ManagemeNt, 52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Parker, S. K., & Grote, G. (2022). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied Psychology, 71*(4), 1171–1204. https://doi.org/10.1111/apps.12241

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., Gerholz, K.-H., & Schlottmann, P. (2025). Reskilling me softly: Perceived changes in students' skilling when using GenAI in Academic Research Projects. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2025.593

Pinski, M., & Benlian, A. (2023). AI literacy—towards measuring human competency in artificial intelligence. In *Proceedings of the 56th Hawaii international conference on system sciences (Ed.)*, HICSS Proceedings 2023.

Qin, J., & D'ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata, 10*(2–3), 188–204. https://doi.org/10.1080/19386389.2010.506379

Rafner, J., Dellermann, D., Hjorth, A., Veraszto, D., Kampf, C., MacKay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: A case for hybrid intelligence. *Morals & Machines, 1*(2), 24–39.

Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *MIS Quarterly, 43*(1), iii–ix.

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: Assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems, 31*(3), 364–387.

Reddy, P., & Sharma, B. (2020). Digital literacy: A review of literature. *International Journal of Technoethics, 11*(2), 65–94.

Reinhardt, U. E. (1992). Reflections on the meaning of efficiency: Can efficiency be separated from equity? *Yale Law & Policy Review, 10*(2), 302–315.

Ridley, M., & Pawlick-Potts, D. (2021). Algorithmic literacy and the role for libraries. *Information Technology and Libraries*, *40*(2). https://doi.org/10.6017/ital.v40i2.12963

Rosenfeld, K. N. (2016). Terms of the digital age: Realities and cultural paradigms. In *Handbook of research on media literacy in the digital age* (pp. 115–144). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-4666-9667-9.ch006

Sanusi, I. T., Olaleye, S. A., Agbo, F. J., & Chiu, T. K. F. (2022a). The role of learners' competencies in artificial intelligence education. *Computers and Education: Artificial Intelligence, 3*, 100098. https://doi.org/10.1016/j.caeai.2022.100098

Sanusi, I. T., Olaleye, S. A., Oyelere, S. S., & Dixon, R. A. (2022b). Investigating learners' competencies for artificial intelligence education in an African K-12 setting. *Computers and Education Open, 3*, Article 100083. https://doi.org/10.1016/j.caeo.2022.100083

Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest editorial: Qualitative studies in information systems: A critical review and some guiding principles. *MIS Quarterly, 37*(4), iii–xviii.

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 1/2). *Journal of the Association for Information Systems, 19*(8), 752–774.

Selber, S., & Selber, S. A. (2004). *Multiliteracies for a digital age*. Southern Illinois University Press.

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

Su, J., & Yang, W. (2023). Artificial intelligence (AI) literacy in early childhood education: An intervention study in Hong Kong. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2217864

Su, J., & Zhong, Y. (2022). Artificial intelligence (AI) in early childhood education: Curriculum design and future directions. *Computers and Education: Artificial Intelligence, 3*, Article 100072. https://doi.org/10.1016/j.caeai.2022.100072

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

Trier, M., Kundisch, D., Beverungen, D., Müller, O., Schryen, G., Mirbabaie, M., & Trang, S. (2023). Digital responsibility. *Business & Information Systems Engineering, 65*(4), 463–474. https://doi.org/10.1007/s12599-023-00822-x

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2023). Responsible AI for digital health: A synthesis and a research agenda. *Information Systems Frontiers, 25*(6), 2139–2157. https://doi.org/10.1007/s10796-021-10146-4

Verma, A., Lamsal, K., & Verma, P. (2022). An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education, 36*(1), 63–73. https://doi.org/10.1177/0950422221990990

Wahl, M. F., & Prause, G. (2013). Toward understanding resources, competencies, and capabilities: Business model generation approach. *Entrepreneurship and Sustainability Issues, 1*(2), 67–80. https://doi.org/10.9770/jesi.2013.1.2(1)

Wang, B., Rau, Pei-Luen Patrick, & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 1–14. https://doi.org/10.1080/0144929X.2022.2072768

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

Weber, P., Pinski, M., & Baum, L. (2023). Toward an Objective Measurement of AI Literacy. In *PACIS 2023 Proceedings*. 60. https://aisel.aisnet.org/pacis2023/60

Wegner, D. M. (1987). Transactive Memory: A Contemporary Analysis of the Group Mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of Group Behavior* (pp. 185–208). Springer New York. https://doi.org/10.1007/978-1-4612-4634-3_9

Yamaguchi, S. (2018). Changes in returns to task-specific skills and gender wage gap. *Journal of Human Resources, 53*(1), 32–70.

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

Zamani, E., & Vannini, S. (20225). Digital policy narratives: addressing grand challenges or exacerbating digital inequalities?. *Information Technology for Development*, 1–20.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nina Passlack Is a doctoral candidate at the Faculty of Information Systems and Applied Computer Sciences at the University of Bamberg, Germany. She holds a Master of Science in International Information Systems Management from the University of Bamberg. Her research primarily explores socio-technical issues, with a particular emphasis on responsibility attribution in human-AI interactions and the competencies necessary for the responsible use of AI.

Teresa Hammerschmidt (neè Heyder) is an affiliate member of the Department of Information Systems and Social Networks at the University of Bamberg. She completed her Ph.D. in Information Systems from the University of Bamberg in 2025. Her research primarily investigates the ethical management of human-AI interaction within organizations, with a focus on sociotechnical system theory

Oliver Posegga Is a professor at the Department of Information Systems and Social Networks at the University of Bamberg, Germany. His research focuses on social networks, digital collaboration, and the sociotechnical dynamics of hybrid media systems as well as the ethical management of AI, AI literacy, and digital responsibility.

Paper VII

Passlack, N., Gerholz, K. H., & Schlottmann, P. (2025). Reskilling Me Softly: Perceived Changes in Students' Skilling When Using GenAI in Academic Research Projects. *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*.

https://hdl.handle.net/10125/109440

Reskilling Me Softly: Perceived Changes in Students' Skilling When Using GenAI in Academic Research Projects

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

Karl-Heinz Gerholz
University of Bamberg, Germany
karl-heinz.gerholz@uni-bamberg.de

Philipp Schlottmann
University of Bamberg, Germany
philipp.schlottmann@uni-bamberg.de

Abstract

The increasing use of Generative Artificial Intelligence (GenAI) in education for academic research projects is reshaping how we acquire information and generate knowledge. Generative AI is defined as computational techniques generating content, such as text from training data, which drives significant skill changes. This raises the question of how GenAI prompts reskilling, involving both deskilling (reduced skill development) and upskilling (acquisition of new skills). The balance between these processes is crucial, as overreliance on GenAI in academic research projects could lead to a performance gap due to, for instance, diminished critical thinking. Semi-structured interviews with students using GenAI for their academic research projects provide insights into shifts in skill requirements and the implications for future academic competence levels. The paper outlines a potential skill gap and concludes that individuals' motivational factors impact reskilling.

Keywords: Generative AI, education, skills, digital literacy, academic research projects

1. Introduction

"The greatest obstacle to discovery is not ignorance - it is the illusion of knowledge" was once remarked by Daniel Boorstin, a highly regarded American intellectual (Krucoff, 1984). Though this observation dates back to 1984, its relevance has only grown in the contemporary context of Generative Artificial Intelligence (GenAI). The latter "refers to computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data" (Feuerriegel et al., 2024, p. 111). The increased usage of GenAI, for example, for academic research projects, is expected to influence how we receive information and generate knowledge (Chen et al., 2023; Shallari & Hussain, 2024). The use of artificial intelligence (AI) may influence how and what we learn and could lead to a shift in skills requirements (Benbya et al., 2021; Rafner et al., 2022). This implies that some skills may become less important through increased GenAI usage over time, while new ones are required (Rafner et al., 2022).

Prior research refers to reskilling, which is the process of individuals acquiring fresh skills and competencies and refining existing ones with new demands due to interacting with AI (Rafner et al., 2022). Reskilling implies two streams of skilling: deskilling and upskilling. Deskilling describes reduced skill development since skills can be displaced to AI (Chuang, 2020) and "the loss of professional skills due to technological or work practice changes" (Rafner et al., 2022, p. 26). Upskilling, by contrast, refers to the acquisition of competencies essential for augmenting both comprehension and proficiency in utilizing AI systems (Jaiswal et al., 2022) to improve efficiency (Moore et al., 2020). When humans and AI fail to engage in "mutual learning", it can result in a "performance gap," limiting the potential outcomes (Rafner et al., 2022, p. 30).

Recent research has been done on increased technology use in higher education (Lee et al., 2024), including AI for academic research projects (Chan & Zhou, 2023; Chen et al., 2023; Lockwood & Castleberry, 2024). Relying more on GenAI can increase efficiency (Chan & Zhou, 2023). However, increased application of GenAI "at the expense of expertise seems a short-sighted solution" (Sutton et al., 2018, p. 17). Deskilling can create a high level of dependency when students do not develop skills to, for instance, critically reflect on the output. We can only manage potential negative consequences of students using GenAI for their academic research projects if we

URI: https://hdl.handle.net/10125/109440
978-0-9981331-8-8
(CC BY-NC-ND 4.0)

understand reskilling. We, therefore, aim to answer the following research question: *What skills are affected by reskilling through increased GenAI use for academic research projects in higher education?*

Accordingly, we conducted semi-structured interviews with higher education students using GenAI for their academic research projects, seeing the need for domain-specific investigations on deskilling (Rafner et al., 2022).

The interviews outlined differences in skills affected by reskilling depending on the type of task in which GenAI is applied. We further provide insights into what skills can be displaced to GenAI without any expected negative long-term consequences and which skills affected by deskilling might negatively affect students' competence levels in academia. We thereby reflect on both perspectives: On the one hand, we highlight the chances of deskilling in case of less relevant skills for future work (such as more time to focus on creativity skill development instead of routine skills). On the other hand, we shed light on unintended deskilling processes as important skills that would be required for responsibly conducting research and academic research projects, which can be referred to as a skill gap (Morandini et al., 2023).

We provide theoretical and practical insights into which skills can be affected by reskilling when using GenAI for academic research projects, thus, presenting three key findings. Firstly, the paper reveals a critical skill gap: students recognize the importance of critical reflection when using GenAI but experience a decline in it due to their reliance on GenAI for increased efficiency. Secondly, despite believing their performance would improve without GenAI, students consistently choose to use it because of its efficiency. Thirdly, motivation affects reskilling: students driven by intrinsic factors fear deskilling with increased GenAI use, while those with extrinsic motivation report rather upskilling.

The paper is structured as follows. We outline the theoretical background in the next section (2). The data collection and analysis approaches are explained in section 3. Section 4, presents the results and discusses them theoretically afterwards (5). Finally, we outline our conclusion (6).

2. Theoretical background

Before we dive deeper into which skills might be affected by reskilling due to the increased use of GenAI to complete academic achievements, we need to understand the differences between deskilling, upskilling, and reskilling (2.1). Next, we present the expectancy theory that can provide possible explanations for processes linked to reskilling that we may observe in the qualitative data. Afterward, we outline why there might be something such as deskilling and upskilling (2.3). Figure 1 summarizes the key dimensions relevant to our study, which are further explained in the following.

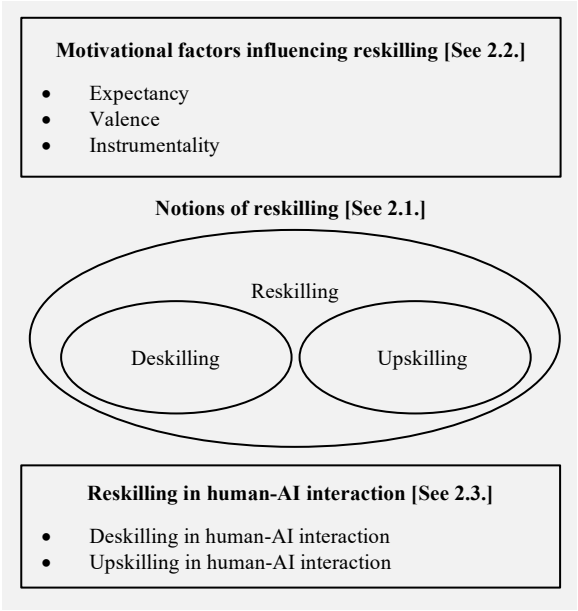


Figure 1. Theoretical framework

2.1. Notions of reskilling: Deskilling and upskilling

The increased use of GenAI can affect individuals' abilities in different ways, which are linked to the notion of reskilling. Reskilling has been defined in preceding studies as the dynamic process whereby individuals acquire new skills and refine existing competencies in response to the evolving demands triggered by the implementation of AI technologies (Rafner et al., 2022). This process encompasses deskilling and upskilling. Deskilling refers to the diminishing emphasis on skill development, often attributable to the displacement of certain tasks by AI (Chuang, 2020), and is characterized by the erosion of professional capabilities resulting from technological advancements or shifts in work practices (Rafner et al., 2022). By contrast, upskilling involves the acquisition of competencies necessary for enhancing the understanding and proficiency in utilizing AI systems (Jaiswal et al., 2022), ultimately aimed at optimizing operational efficiency (Moore et al., 2020). Failure to facilitate "mutual learning between humans and AI" may lead to a "performance gap" (Rafner et al., 2022, p. 30).

2.2. Expectancy theory: motivational factors that may influence reskilling

Vroom's expectancy theory (1964) helps us to understand the processes of deskilling and upskilling better, as motivation is key to continuous learning (Gopalan et al., 2017). The theory suggests three key aspects affecting individuals' motivation: expectancy, valence and instrumentality.

i. **Expectancy** suggests that individuals are willing to put more effort into a task if they expect that effort to influence their performance positively. In this study context, we are interested in finding to what extent students believe that their skills and commitment are decisive for the success of their academic research projects.
ii. **Valence** considers "the importance, attractiveness, desirability, or anticipated satisfaction with outcomes" (Van Eerde & Thierry, 1996, p. 576). It can be referred to the extent to which students are interested in very good results for their research projects.
iii. **Instrumentality** "has been interpreted not only as a relationship between an outcome and another outcome but also as a probability to obtain an outcome" (Van Eerde & Thierry, 1996, p. 576). Concerning this, we are interested in the extent to which students believe that GenAI tools can improve their performance in research projects.

As outlined, we assume that all aspects drive individuals' motivation (HemaMalini & Washington, 2014) and, consequently, reskilling as learning new things is highly linked to motivational factors (Gopalan et al., 2017). This knowledge will help us to find causes for deskilling and possible measures to avoid the deskilling of skills still relevant to conducting academic research projects, although using GenAI in the future.

2.3. Reskilling in human-AI interaction

While prior research on human-AI interaction addresses reskilling, for instance, in the context of medical professionals (Rafner et al., 2022), qualitative investigations on reskilling in the context of GenAI usage for academic research projects are needed. Yet, prior research in AI contexts will help us to interpret our new findings. Therefore, we will highlight the key findings of previous research on reskilling in human-AI interaction in the following.

2.3.1 Deskilling. Upon screening the current literature, we found different skills to be affected by the deskilling process resulting from reliance on AI in human-AI interaction.

Increased reliance on AI can have certain impacts on individuals' perceived autonomy. Support in decision-making based on AI was found to make individuals attribute autonomy to the AI system (Mayer et al., 2021), implying they decreasingly assume responsibility. Furthermore, a (perceived) loss of autonomy can have severe impacts when the AI support system has technological issues. Employees may also be unable to make decisions independently when using the system from day one of a new job (Strich et al., 2021). In addition, when relying on AI-based advice, individuals may also not understand how to interpret the output, which can result in decreased skill development and affect employees' domain knowledge (Rafner et al., 2022) or creativity (Zirar et al., 2023).

2.3.2 Upskilling. Besides the effects of deskilling, using GenAI may also lead to the development of new skills. Upskilling focuses on skills that are essential for the use of GenAI or emerge through interactions with technology (Rafner et al., 2022). In this context, current research particularly emphasizes transversal skills that allow individuals to apply them in different contexts (Hart et al., 2022). Morandini et al. refer to "critical thinking, problem, solving, communication, and collaboration" as typical critical transversal skills as "they enable workers to adapt to new technologies and processes and to continuously learn and develop in the face of rapidly changing technology" (2023, p. 46).

Other examples are English fluency, the language of AI documentation and interfaces, that empowers individuals to navigate better and utilize GenAI technologies. This proficiency fosters a deeper understanding of AI tools, enhancing their use and innovation (Irawan et al., 2022). In addition to language skills, basic mathematical concepts are proposed to enable effective AI applications (Verma et al., 2022). Analytical and critical thinking skills are also suggested to be vital for interpreting AI's output and arriving at well-informed conclusions (Delanoy & Kasztelnik, 2020).

Given that current research shows that the use of GenAI significantly impacts students' writing processes concerning structuring academic essays (Guo et al., 2023), sharpening the argumentation (Storey, 2023), enhancing text quality, or analyzing data (Malik et al., 2023), we aim to investigate which skills might be affected by reskilling in academic research projects.

The theoretical background and initial findings of prior research help us to analyze our explorative study, aiming to outline shifts in skills caused by GenAI use for academic research projects in higher education. The next sections present our methods of data collection (3.1) and data analysis (3.2) processes.

3. Methods

We interviewed 12 students, which allowed us to compare differences in reskilling processes in the context of GenAI use in academic research projects, depending on different tasks and subjective experiences. The data collection relies on semi-structured interviews with students who have used GenAI for academic research projects for more than six months.

3.1. Data collection: Semi-structured interviews

We interviewed 12 students to identify the skills affected by reskilling when using GenAI in doing research. The students were recruited from a master's course at a German University. All students reported using GenAI for their academic research projects and had already studied for over six semesters at a university. We chose only students who had studied for more than six semesters as they have had experience writing academic research papers without GenAI tools (in their bachelor courses) and, thus, can compare it with the new setting. All interviews were held in German and translated into English afterward. The students received five credit points for their exam performance for participating in the interviews. The interviews took place from May 20 to 31, 2024, and each lasted about 20 minutes.

Before starting the interviews, students were asked whether they agreed to be recorded and have their recordings transcribed. Next, we asked them about their course of study, which semester they studied and their understanding of GenAI.

We then asked the following main questions:

- For which tasks do you use GenAI tools when doing your academic research project?
- Why do you use GenAI tools for doing your research project?
- Which new skills have you acquired through using GenAI tools for your academic research project? [*upskilling*]
- Which skills do you require to use GenAI for your academic research projects? [*reskilling*]
- To what extent do you believe you lost or did not learn certain skills using GenAI tools for your academic research projects? [*deskilling*]
- To what extent do you believe that your skills and commitment are decisive for the success of your academic research projects? [*expectancy*]
- To what extent do you think GenAI tools can improve your performance in your academic research project? And why do you think so? [*instrumentality*]
- How important is a very good result for your academic research project? Please explain what your main motivators are. [*valence*]

The demographic data of the 12 students interviewed are shown in Table 1 and are numbered to ensure anonymity.

Table 1. Demographic data of participants

Reference	Sex	Semester	Masters Course
I1	m	1st	Business Education II
I2	w	1st	Business Education II
I3	m	3rd	Business Education II
I4	w	1st	Business Education II
I5	m	1st	Business Education II
I6	m	1st	Business Education II
I7	w	1st	Business Education I
I8	m	1st	Business Education II
I9	w	1st	Business Education II
I10	w	1st	Communication
I11	w	1st	Business Education II
I12	w	1st	Business Education I

3.2. Data analysis: Open, axial and selective coding

The data analysis followed the approach suggested by Gioia et al. (2013). Consequently, our coding schema (see Figure 2) comprises 1st order concepts, 2nd order themes and aggregated dimensions. The 1st order constructs were derived from the data itself. Hence, we stuck to the wording of our participants. In the second step, we derived the 2nd order themes that cluster the 1st order concepts according to themes. In doing so, we interpreted the 1st order constructs based on our theoretical background. In the last step, we assigned the 2nd order themes to the three skilling processes (see Section 2.1.) and the three motivational factors (see Section 2.2) as aggregated dimensions. The coding process, conducted by two authors, followed an iterative structure. The coding scheme was refined through multiple discussion rounds until both coders reached a consensus.

Figure 2 presents a snippet of our coding schema with the 1st order concepts, the 2nd order themes and the aggregated dimensions.

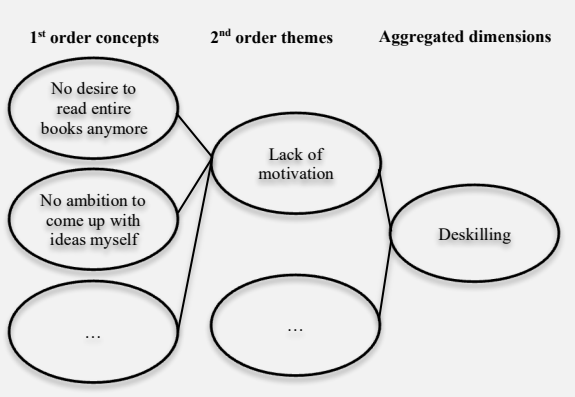


Figure 2. A snippet of the coding schema

4. Findings: Reskilling through GenAI usage in academic research projects and students' motivation

The following pages first present the various tasks where students utilize GenAI in their academic research projects (see 4.1). Subsequently, the study's results are outlined in relation to the skilling processes: upskilling (see 4.2), deskilling (see. 4.3) and reskilling (see 4.4). Table 2 presents an overview of the main findings. Finally, we outline what motivates students in the context of their academic research projects (see 4.5) to better understand how this may affect their self-perception of their personal reskilling processes.

4.1. Students reported that GenAI helps them primarily with brainstorming, summarizing, paraphrasing, understanding and analyzing data, as well as emotional support.

Many students (n = 6) stated that they use GenAI to summarize papers. As one student explained: "*[This way] I can simply check whether the paper is relevant to me. I don't have to read through it first and search to see if anything is interesting, but I can get a summary*." (I9)

Six students reported that they use GenAI mainly for brainstorming and generating new ideas to "*just get inspiration so that [they] don't just sit in front of a blank piece of paper and think, what should I do now?*" (I9) Another student described her process of using GenAI: "*When we get a work assignment and I just don't know where to start, how to start, so that I use it [ChatGPT] like a collection of ideas … I put a question in there … and then it gives me an overview, a rough concept … and then I know how I could start.*" (I7)

Table 2. List of skills affected when using GenAI in academic research projects

Skills affected		Reference
DE-skilling	in-depth understanding	I2, I4, I8, I9
	text comprehension	I3, I9
	self-motivation	I8, I11
	critical reflection	I10, I11, I12
	information search	I1, I3, I5, I6, I7, I12
	creativity and independent thinking	I1, I6, I7, I9, I10, I11, I12
	attention span and concentration	I7, I8, I10, I12
	writing and communication skills	I3, I5, I6
UP-skilling	logic reasoning	I8
	domain knowledge	I3, I4, I7, I8
	precise communication	I3, I5, I7, I9, I11
	digital expertise	I7
RE-skilling	precise communication and prompting	I1, I8, I10, I11
	self-motivation to learn	I7, I12
	digital skills	I3, I4, I5, I7, I9
	critical reflection	I1, I2, I3, I4, I5, I6, I7, I8, I9, I10, I11, I12

Additionally, some students (n = 6) reported using GenAI to paraphrase their own words to make them sound smoother or reframe existing research to avoid direct quotes. As one student framed it: "*When I say 'Okay, the text fits what I want to say, but I'd like synonyms or I'd prefer to rearrange the sentence so I don't have to quote it verbatim,' I simply use it to avoid the cognitive effort. For me, it's just a more efficient way of working.*" (I9)

Moreover, students use GenAI for research and data analysis in particular because "*it saves a lot of time, of course. So the advantage is that you can search through as much data in just a few seconds as you wouldn't be able to do yourself. That's the biggest advantage, I would say*." (I12)

Interestingly, one student reported how she seeks to receive emotional support from ChatGPT when she is uncertain about the quality of her own work: "*What I actually did for the very first time was, after a small emotional moment regarding my bachelor's thesis*

when I wasn't quite sure if I was doing everything right or started questioning everything a bit, I sent ChatGPT all the data I had collected so far, my research question, and my analysis. I asked, 'Do you think this makes sense? Do you think what I'm doing is coherent?' and I also asked it to be nice in its response." (I10) In doing so, she hoped it would provide her with some sort of emotional support. She added: "*I know it's not the same as asking my supervisor, but at that moment, I was really just looking for a bit of reassurance, a small safety net.*" (I10)

4.2. Students reported increased communication skills, digital competencies and broader domain knowledge.

Starting with the perceptions of upskilling, four of the students said that using GenAI helps them to understand domain-specific problems better and, thus, supports their learning process of domain knowledge when writing academic essays. This can be seen, for example, in the use of statistics software. As one student stated: *"As far as coding in R is concerned, I [...] asked ChatGPT in this case and ChatGPT then also explained virtually what code you ultimately need, but also how to get there*." (I8) Another student pointed out that she learns English while translating with GenAI tools and *"because I [the student] translate a lot of things, I have to say that some things [vocabulary] still stick*" (I7).

The skills most reported to be part of upskilling (n = 5) were communication skills and the ability to write precise formulations. Using precise commands when working with GenAI tools was said to be essential for good results from the students' perspective. This is underlined by statements such as "*and so you learn how to put the most important things succinctly but qualitatively in terms of content"* (I3) or "*[…] at the beginning it was more the case that I used keywords, but now you can really say exactly what you want and the more you write about it and simulate it [ChatGPT], the better the results*" (I9). One student mentioned that her logical reasoning was improving and another said that her digital skills were also developing. She said: "*[If] you use digital tools, you also need digital skills to understand the whole thing and use it properly*" (I7).

4.3. Students reported decreased creativity, critical reflection, learning abilities and concentration.

Regarding deskilling, the majority of students (n = 7) reported that their creativity and independent thinking might get lost through increased GenAI use for academic research projects. One student (I11) explained that she uses ChatGPT for brainstorming and said, "*That's how you unlearn this creative thinking a bit*" (I11).

Furthermore, GenAI use was also reported to have certain effects on students' learning skills. Some students reported that using GenAI for their research projects negatively affects their motivation to do it themselves (n = 2). One student said: "*I've noticed myself that when I have a task where I have to read a long text, I think to myself, oh no, there's far too much text. I want to enter it into ChatGPT right now and have it summarized.*" (I11) Other students explained how GenAI decreases their critical reflection skills (n = 3). As one student framed it: "*And because I rely so much on that, what [ChatGPT] paraphrases is correct, I notice that I have become a little less precise in double checking.*" (I10).

Additionally, some students reported that using GenAI to get explanations on different topics results in a poorer in-depth understanding of the topic than conducting thorough research or reading a book rather than relying on a short summary generated by a GenAI since "*then the anchoring of knowledge is missing*" (I2). Moreover, some students (n = 4) noted that they felt their attention span and independent thinking had decreased since they regularly used ChatGPT for their academic projects.

Skills linked to language were also reported to be affected by deskilling through increased GenAI use. One student who also worked as a teacher shared her concerns: "*I can see how difficult it is for children to understand a text. And if they just put it [a text] into AI [to get a summary], so to speak, then they really don't have any understanding of the text at all.*" (I9)

Similarly, some students explained that they perceived GenAI usage negatively affecting their text comprehension (n = 2) and writing and communication skills (n = 3).

4.4. Students reported precise communication, self-motivation to learn and critical reflection to be critical for GenAI use.

In terms of reskilling, students (n = 11) emphasized that critical reflection is necessary for the proper use of GenAI in academic research projects. It is important *"that you can also reflect critically [...] because there are often things that just don't fit*" (I11) and "*then you should definitely question the results and if suitable results are offered, you should, of course, read them again in detail*" (I3).
In addition, the ability to formulate precisely and write confusion-free prompts is mentioned by four students. The students' statements show that this is a matter of practice and can be learned. As one student explained: "*And it actually took me a while until I had the ability to formulate the prompts in such a way that I knew in advance that I was going in the right direction.*" (I8)

Other students (n = 4) mentioned that it is important to develop digital skills in order to be able to use GenAI in a targeted manner. This applies "*[to] digital skills, just that you know that the tools are available, that you have a rough overview, maybe what you can do with them*" (I7*).* It is emphasized that it is necessary to be able to select suitable tools. Furthermore, in order to learn how to use the tool, self-motivation was reported to be relevant to "*learning by doing. So you write in and then maybe ask yourself what else can the tool do*" (I12).

4.5. Students' motivation for their academic research projects was mainly driven by the desire to receive good grades, prove something to themselves and be efficient.

Concerning students' perceived *valence* of receiving good grades, most students stressed that they are very motivated to perform well, yet the driving factors for their motivation differed. Some students (n = 6) stated that achieving good grades is important for their later career development. As one student framed it: "*Especially when you go into the private sector, your grades play a role, and the better you are, the easier it is for you*" (I8). By contrast, other students (n = 6) are primarily intrinsically motivated and want to prove their self-worth and learn something (n = 1). As one student framed it: "*I'm just doing it for myself because I've set myself a personal goal*." (I9)

Regarding the motivation for using GenAI, which we linked to *instrumentality*, all students emphasized the timesaving and efficiency of GenAI when using it for academic research projects. As one student framed it: "*Well, I think I would simply invest more time, effort and, energy into it [the research project] than I would now with an AI*" (I10). By contrast, other students (n = 4) stated that they perceive to have more effort when using GenAI through double-checking outputs. However, GenAI was reported as being critical for the project's success (n = 5) as "*an incredible amount of data can be processed; I'm not capable of that*." (I5). Interestingly, other students (n = 2) reported that they would often not understand certain connections in the literature without ChatGPT, which would minimize their performance. One explained: "*If I haven't fully understood the theory behind it ... ChatGPT can explain it to me in a simpler way than it's presented in academic literature.*" (I2)

Concerning students' *expectancy* that their effort is decisive for their project's success, all stated their personal effort, skills and commitment to be critical for good performance in the research projects. The GenAI, for instance, was reported to lack creativity (n = 3) and GenAI hallucinations were mentioned as a reason (n = 3) why students think their personal skills, such as critical reflection, would be crucial for good grades: "*When I type, 'give me 10 sources on the topic of knowledge management in companies' or something like that, I do get sources, but the books don't even exist. So, it's kind of, well, worthless if it's not theoretically sound*." (I9). Another student mentioned that human skills are critical for good research since only humans "*can place connections and relationships with other topics that should be critically examined into context, meaning it perceives the overall context*." (I2)

5. Discussion and implications

This study's findings aim to shed light on what skills might be affected by reskilling through increased GenAI use for academic research projects in higher education and what may motivate students to rely on their own skills for doing academic research projects. Based on semi-structured interviews with students who use GenAI for their academic research projects, the study provides an overview of several skills that become less relevant or, by contrast, more crucial when using GenAI in academia. The following section outlines three key findings that contribute to IS research and require further investigation.

Firstly, the paper indicates a critical skill gap since, although students recognize the importance of critical reflection when using GenAI for academic research projects, they actually experience a decline in critical reflection by relying heavily on GenAI as they perceive it to increase their efficiency in academic research projects (linked to *instrumentality*). Similarly, Malik et al. (2023) address issues linked to

overreliance and concerns that AI use might lower critical thinking abilities. Further research would be required investigating the long-term implications of GenAI use and academic integrity.

Secondly, the findings suggest that despite some students believing that their performance would improve if they worked independently without using GenAI (linked to *expectancy*), they consistently choose to use AI due to higher efficiency. While Nguyen et al. (2024) found that high-achieving students use GenAI more frequently and more consciously than low-achieving students, who only use such tools as support, future research is needed in this regard.

Thirdly, the findings suggest that individuals' motivation for good performance in research projects (linked to *valence*) may affect reskilling. Students motivated by external factors, such as better opportunities on the labor market, reported more skills that referred to upskilling when speaking about their GenAI use. By contrast, those who perceived intrinsic motivation for performing well in their research projects, such as a desire to learn, reported more that they sometimes fear downskilling through increased GenAI use. Future research would be welcomed that quantitatively investigates the relationships between motivational factors and perceived reskilling through GenAI use.

In addition to the findings outlined above, the study has several limitations. While we focused on master's students in business education, including other stakeholders, such as PhD students and academic researchers who use GenAI in their research projects, could provide additional insights. Furthermore, the findings cannot be generalized; larger sample sizes could enhance the significance. Moreover, the answers can be biased by social desirability, in cases where the students do not want to disclose their usage of GenAI in their academic research projects. Future research is also encouraged to quantitatively investigate the extent to which our findings can be validated and expanded.

6. Concluding comments

The introductory quote stating that "the greatest obstacle to discovery is not ignorance - it is the illusion of knowledge" underscores the importance of critical reflection in the context of GenAI in academic research. As our study highlights, reliance on GenAI can create an illusion of proficiency while leading to a decline in critical thinking skills. Therefore, we must recognize and address this skill gap to ensure that GenAI usage in academic research projects enhances rather than hinders the development of relevant skills. This paper outlines how students' motivation might affect their perceived deskilling and upskilling processes. Future research and educational practitioners are encouraged to foster critical reflection on potential unintended consequences of GenAI use for academic research projects.

7. References

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special issue editorial: Implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281-303.

Chan, C. K. Y., & Zhou, W. (2023). Deconstructing student perceptions of generative AI (GenAI) through an expectancy value theory (EVT)-based instrument (arXiv:2305.01186). arXiv.

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, *25*(1), 161–182.

Chuang, S. (2020). An empirical study of displaceable job skills in the age of robots. *European Journal of Training and Development*, *45*(6/7), 617–632.

Delanoy, N., & Kasztelnik, K. (2020). Business open big data analytics to support innovative leadership decision in Canada.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126.

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.

Gopalan, V., Bakar, J. A. A., Zulkifli, A. N., Alwi, A., & Mat, R. C. (2017). A review of the motivation theories in learning. *Aip Conference Proceedings*, *1891*(1).

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection (arXiv:2301.07597). arXiv.

Hart, J., Noack, M., Plainmauer, C., & Bjørnåvold. (2022, May 17). Towards a structured and consistent terminology on transversal skills and competences. *Brüssel: Europäische Kommission und Cedefop. Towards a structured and consistent terminology on transversal skills and competences Esco (europa. eu*, 1, 645-670.

HemaMalini, P. H., & Washington, A. (2014). Employees' motivation and valued rewards as a key to effective QWL-from the perspective of expectancy theory. *TSM Business Review*, *2*(2), 45.

Irawan, C. B., Sasongko, B., & Afriani, I. (2022). The role of English and technology on human performance in international business. *International Journal of Applied Business and International Management*, *7*(2), 30–40.

Jaiswal, A., Arun, C. J., & Varma, A. (2022). Rebooting employees: Upskilling for artificial intelligence in

multinational corporations. *The International Journal of Human Resource Management*, *33*(6), 1179–1208.

Krucoff, C. (1984). "The Washington Post" The 6 O'Clock Scholar. https://www.washingtonpost.com/archive/lifestyle/1984/01/29/the-6-oclock-scholar/eed58de4-2dcb-47d2-8947-b0817a18d8fe/

Lee, S., Jang, W. W., & Rollins, M. (2024). Using Minecraft Education Edition to enhance 21st century skills in the college classroom: A mixed methods study. *Proceedings of the 57th Hawaii International Conference on System Sciences*.

Lockwood, A. B., & Castleberry, J. (2024). Examining the capabilities of GPT-4 to write an APA-style school psychology paper. *Contemporary School Psychology*, 1-8.

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, *5*, 100296.

Mayer, A.-S., Haimerl, A., Strich, F., & Fiedler, M. (2021). How corporations encourage the implementation of AI ethics. *ECIS 2021 Research Papers*.

Moore, D., Haines, K., Drudik, J., Arter, Z., & Foley, S. (2020). Upskill/backfill model of career pathways advancement: The Nebraska Vocational Rehabilitation Approach. *Journal of Applied Rehabilitation Counseling*, *51*(3), 208–221.

Morandini, S., Fraboni, F., De Angelis, M., Puzzo, G., Giusino, D., & Pietrantoni, L. (2023). The impact of artificial intelligence on workers' skills: Upskilling and reskilling in organisations. *Informing Science: The International Journal of an Emerging Transdiscipline*, *26*, 39–68.

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864.

Rafner, J., Dellermann, D., Hjorth, A., Veraszto, D., Kampf, C., MacKay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: A case for hybrid intelligence. *Morals & Machines*, *1*(2), 24–39.

Shallari, I., & Hussain, M. (2024). Assignments in the ChatGPT-era: Case study on plagiarism in digital systems design courses. *Proceedings of the 57th Hawaii International Conference on System Sciences.*

Storey, V. A. (2023). AI technology and academic writing: Knowing and mastering the "craft skills." *International Journal of Adult Education and Technology (IJAET)*, *14*(1), 1–15.

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*, *22*(2).

Sutton, S. G., Arnold, V., & Holt, M. (2018). How much automation is too much? Keeping the human relevant in knowledge work. *Journal of Emerging Technologies in Accounting*, *15*(2), 15–25.

Van Eerde, W., & Thierry, H. (1996). Vroom's expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology*, *81*(5), 575–586.

Verma, A., Lamsal, K., & Verma, P. (2022). An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education*, *36*(1), 63–73.

Vroom, V. H. (1964). *Work and motivation*. Wiley.

Zirar, A., Ali, S. I., & Islam, N. (2023). Worker and workplace Artificial Intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, *124*, 102747.

Paper VIII

Leffrang, D., Passlack, N., Müller, O., & Posegga, O. (2026). Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice. *Business and Information Systems Engineering (BISE)*.

https://doi.org/10.1007/s12599-026-01000-5

Bus Inf Syst Eng
https://doi.org/10.1007/s12599-026-01000-5

RESEARCH PAPER

Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice

Dirk Leffrang · Nina Passlack · Oliver Müller · Oliver Posegga

Received: 14 August 2023 / Accepted: 22 December 2025
© The Author(s) 2026

Abstract Generative Artificial Intelligence (GenAI) systems, such as large language models (LLMs), are increasingly permeating everyday tasks; yet, they sometimes present false or misleading information as fact – a phenomenon known as hallucination. As GenAI becomes more accessible to the public, it is essential to understand how users' AI literacy shapes their reliance on imperfect advice from AI systems. Building on the concept of correspondence bias, the study examines how individuals with different levels of AI literacy respond to faulty GenAI advice. Evidence from an online programming experiment with 542 U.S. programmers shows that individuals with higher AI literacy rely less on GenAI advice, particularly when the advice is flawed. Correspondence bias provides a plausible explanatory mechanism for these findings and helps reconcile mixed results in prior research on AI literacy. Overall, the findings offer a more nuanced perspective on the benefits and risks of AI-literacy-driven mistrust. This informs education, integration, and evaluation initiatives for GenAI while cautioning against naive evaluation strategies.

Keywords AI literacy · Artificial intelligence · AI education · Algorithm aversion · Laboratory experiment

Accepted after two revisions by Alexander Benlian.

D. Leffrang (✉) · O. Müller
Paderborn University, Paderborn, Germany
e-mail: dirk.leffrang@uni-paderborn.de

N. Passlack · O. Posegga
University of Bamberg, Bamberg, Germany

1 Introduction

Advances in Artificial Intelligence (AI), particularly Generative AI (GenAI), are fueling its increasing use among knowledge workers such as programmers (e.g., Surameery and Shakor 2023; Dell'Acqua et al. 2023). GenAI refers to AI systems that create new content, such as text, images, or code, by learning from existing data (Openai 2024). However, these systems are not flawless and can produce misleading or false information, known as "hallucinations" (Maynez et al. 2020). When encountering such "bad advice," users may attribute it to situational factors (e.g., insufficient context in the prompt) or dispositional factors of the GenAI (e.g., model limitations). Such attributions are crucial as they influence advice-taking both while receiving bad advice and in future tasks.

AI-based systems have become increasingly user-friendly, making them accessible not only to experts but also to individuals from diverse backgrounds (Mädche et al. 2019). This shift creates a need for new competencies to engage effectively with AI, commonly referred to as AI literacy (Long and Magerko 2020). AI literacy refers to 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). Enhancing individuals' AI literacy enables them to align their evaluation of AI's appropriateness with their delegation of decisions to AI (Pinski et al. 2023). For instance, understanding the rationales behind AI output as one competency linked to AI literacy is crucial for appropriate interaction and dealing with the inscrutability of AI (Asatiani et al. 2021; Lebovitz et al. 2021). Therefore, gaining a deeper understanding of individuals' interactions with AI-based tools is imperative to mitigate potential negative

Published online: 12 May 2026

outcomes in case of bad advice from an AI (Benbya et al. 2021).

However, there are mixed results regarding the direct effects of AI literacy on advice-taking, while the effects of AI literacy remain among the least examined aspects of the AI literacy literature (Pinski and Benlian 2024a). Prior studies suggest that increasing AI knowledge, as one component of AI literacy, can foster greater reliance on AI by raising delegation rates (Pinski et al. 2023) and advice-taking (Panigutti et al. 2022). Yet these studies primarily evaluated AI on unbalanced datasets dominated by good advice (e.g., Tully et al. 2025). By contrast, individuals with greater statistical knowledge tended to trust algorithms less (Marmolejo-Ramos et al. 2025). Other studies found that individuals with a limited understanding of AI are more likely to overestimate the intelligence of AI (Ehsan et al. 2021) due to unrealistic expectations (Dietvorst et al. 2015), which can lead to increased advice-taking. Thus, we aim to investigate how AI literacy influences advice-taking by introducing a novel perspective that examines attributional reasoning, specifically through the lens of correspondence bias. We position ourselves within this debate and argue that when advice quality is evenly split between good and bad quality advice, AI literacy has a negative correlation with advice-taking, as higher literacy mitigates correspondence bias and fosters more cautious and selective engagement with GenAI.

While expectations in GenAI are growing, particularly for enhancing productivity in domains such as programming (Peng et al. 2023), AI is not immune to errors and can provide misleading advice (Lebovitz et al. 2021). Detecting and fixing bad advice from GenAI, such as hallucinations, remains an ongoing research frontier (Berente et al. 2021; Ji et al. 2023). We are especially interested in advice-taking while receiving bad advice (task t). Blindly following advice can be problematic if an AI's advice fails to meet the users' expectations (Stieglitz et al. 2023). As with any setback, individuals naturally seek to understand the reasons behind the event (Heider 1958). Bad advice attributed to the advisor's competencies can lead to a rapid decline in the advisor's reputation and thereby decrease future advice-taking (Yaniv and Kleinberger 2000). Therefore, we are also interested in how individuals' AI literacy influences future advice-taking (task t+1) due to different reasoning and causal attributions of a GenAI's hallucination. Thus, we aim to answer the following research question:

What is the influence of AI literacy on individuals' advice-taking behavior while receiving bad GenAI advice and their advice-taking behavior in future tasks?

To address our research question, we build on correspondence bias theory, suggesting that individuals tend to attribute bad advice to dispositional factors (e.g., GenAI's model) rather than situational factors (e.g., forces external to the GenAI, such as a prompt missing context details) (Gilbert and Malone 1995). We conducted an online programming experiment and complemented the quantitative results with qualitative insights on the correspondence bias, thereby adopting a mixed-methods design (Reis et al. 2022). First, we investigated individuals' advice-taking while receiving bad advice from GenAI (task t) and how they take advice in the next task when receiving new advice (task t+1) using the Judge Advisor System (JAS). Therefore, we investigated two factors: the quality of advice (good vs. bad) and individuals' level of AI literacy. Second, we qualitatively analyzed open-ended questions to explore the reasons for differences in advice-taking across AI literacy levels. We found that receiving bad advice from GenAI reduces both the likelihood of taking that advice and the likelihood of taking new advice from GenAI in the future. In addition, we observed a negative correlation between individuals' AI literacy and their tendency to take advice from GenAI. We also found that AI literacy moderated the decline in advice taking triggered by receiving bad advice, such that more AI-literate individuals took less advice when exposed to poor recommendations. However, we could not confirm that more AI-literate individuals, who have previously received bad advice, would take relatively more future good advice from GenAI than less literate individuals. Furthermore, we present three propositions based on our qualitative findings that offer insights into how AI literacy may influence correspondence bias and contribute to an understanding of individuals' decision-making processes, specifically in the context of advice-taking behavior.

We make three theoretical and practical contributions. First, we provide causal evidence on how bad advice and AI literacy shape human reliance on generative AI in a complex programming task, extending prior research that has largely examined simpler predictive systems with numeric outputs. By investigating advice-taking in the context of open-ended, plausibly sounding but variably correct GenAI code suggestions, we broaden empirical understanding to a qualitatively different class of AI systems. Second, we advance theory by identifying AI literacy as a key moderator of attribution processes in human–AI interaction: drawing on correspondence bias theory, we show that users often infer stable flaws in the system from isolated errors, yet individuals with higher AI literacy are more likely to attribute mistakes to situational factors such as missing context, leading to a form of beneficial mistrust that contrasts with the more global negative reactions of low-literate users. These insights refine existing theories of trust and reliance on AI by demonstrating that reactions to AI errors depend not only on performance feedback but also on literacy-dependent cognitive interpretations of AI

behavior. Third, our findings offer actionable implications for designing GenAI systems and interfaces–such as highlighting situational uncertainties, supporting expectation calibration, and providing trust-repair mechanisms–to foster calibrated reliance and more resilient human–AI collaboration.

2 Theoretical Background: The JAS and The Correspondence Bias

To evaluate the extent to which individuals incorporate guidance from GenAI into their own decision-making process and their rationale behind it, we utilize the Judge-Advisor System (JAS) as a framework. In the JAS, a person referred to as a "judge" faces a decision task and forms an initial judgment. Subsequently, the judge receives advice from an advisor, such as a GenAI. The judge then combines their initial judgment with the advice received from the advisor to arrive at a final judgment (Bonaccio and Dalal 2006).

Judging the advice in order to arrive at a final decision can be subject to certain biases. Because judges evaluate both the advisor and the advice, attributional tendencies may influence their reasoning. One such tendency is the correspondence bias, which refers to the tendency of humans to attribute observed behavior to dispositional factors (i.e., reliability, ability, or skills) while underestimating the role of situational influences, such as the environment or situation (Gilbert and Malone 1995; Edwards and Edwards 2022). Dual-process theory offers a valuable lens for understanding this phenomenon. It suggests that individuals initially engage in fast, intuitive processing (System 1), which favors dispositional explanations, while the correction toward situational attributions requires slower, more effortful reasoning (System 2) (Kahneman 2011). The correspondence bias is part of attribution theory, which seeks to explain how individuals interpret the reasons behind specific observations, categorizing them as either situational factors or dispositional factors (Kelley and Michela 1980).

The perception of the situation, behavioral expectations, and the perception of the observed behavior influence individuals' attribution processes (Gilbert and Malone 1995; Edwards and Edwards 2022). Correspondence bias occurs because individuals rely on their causal theories about how situations influence behavior when making judgments about others. In practice, this means correspondence bias persists because people often stop at the intuitive inference and fail to fully engage in the deliberate correction process, especially under cognitive load or when being under time pressure (Kahneman 2011). Thus, individuals tend to attribute behavior to a person's characteristics or dispositional factors, as it is easier and faster for them to access than to evaluate situational dimensions (Gilbert and Malone 1995).

Although researchers first observed correspondence bias in human-to-human interactions, there is a growing tendency for individuals to perceive AI not merely as a tool but as a collaborative teammate (O'Neill et al. 2022), often forming emotional bonds in the process (Müller 2021). Hence, preliminary research has begun to extend the application of correspondence bias to the context of human-AI interactions (Edwards and Edwards 2022), as individuals increasingly respond to AI and humans in similar ways by attributing responsibilities to AI (Hong et al. 2020).

Figure 1 illustrates how insufficient context can create hallucinations, leading to bad advice. It also demonstrates how correspondence bias can influence individuals' attribution of responsibility throughout this process. On a technical level, GenAI tools supporting coding processes, such as GitHub Copilot, access the context of a programming task through code and data shared with the GenAI (e.g., Salva 2024). However, the prompt given to the GenAI (i.e., code written by the user) may lack contextual information (i.e., data columns of a file not shared with the GenAI). Consequently, GenAI relies on its inherent model to provide coding advice and suggest the most likely column names based on the available code. These predicted column names may not exist in the actual data, leading to hallucinations and potentially providing incorrect advice. Therefore, in the context of this study, situational factors (such as the prompt lacking contextual details) can entirely explain the poor advice, as the inherent model of the GenAI correctly provided the most likely suggestion, even though it was incorrect in this situation.

The programmer (judge), on the receiving end of GenAI's advice, can only observe its performance and seek to identify potential causes and explanations for instances of bad advice. Those explanations are likely to shape the programmer's behavior while receiving bad advice (task t) and in future advice-taking scenarios (task t+1). Prior research suggests that the correspondence bias can shape such attribution processes, leading individuals to overemphasize dispositional factors (e.g., the quality of the GenAI model) as the primary cause of the observed behavior (e.g., bad advice) while underestimating situational factors (e.g., contextual limitations in the prompt) that may have also influenced the outcome.

Following Gilbert and Malone (1995), three established explanations contribute to the phenomenon of correspondence bias, as illustrated on the right side of Fig. 1: lack of awareness of the situation (i); unrealistic expectations of behavior (ii); and incomplete corrections of inherent inferences (iii). Those three explanations will guide the

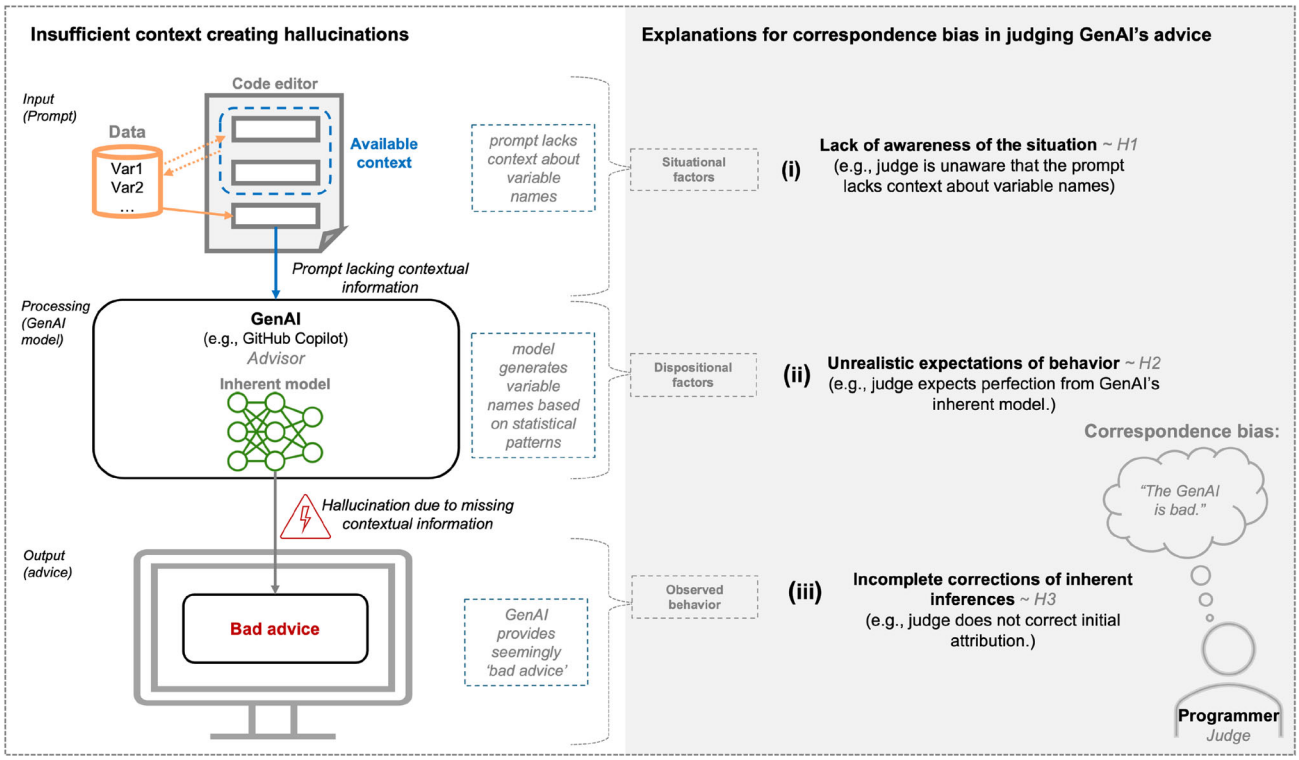


Fig. 1 Insufficient context creates hallucinations and may cause correspondence bias

development of the hypothesis (see Sect. 3) and the derivation of the qualitative questions (see Sect. 4.4). These explanations provide insights into and explanations for differences in advice-taking behavior among individuals possessing varying levels of AI literacy, as we expect higher levels of AI literacy to mitigate the correspondence bias. We outline the psychological foundation of the correspondence bias in the following:

(i) *Lack of awareness of the situation* refers to individuals overlooking contextual circumstances influencing a GenAI's advice due to an incomplete understanding of how such circumstances affect the GenAI's functioning. While situational circumstances, such as data columns in the file being inaccessible to GenAI, significantly influence GenAI's advice, attributing bad advice to situational factors requires an awareness of their existence. Previous studies have noted that when judges are uncertain about the reasons behind an advisor's behavior, they often attribute it to dispositional factors of the advisor (Gilbert and Malone 1995). For instance, organizational employees attributed the cause of critical incidents in human-computer interactions to failures of advisors, such as software, in 74% of cases (Ortiz de Guinea 2016). In the context of GenAI, situational factors such as missing context details in the prompt can cause critical incidents, potentially leading to hallucinations.

(ii) *Unrealistic expectations of behavior* implies that judges can experience correspondence bias if a GenAI's behavior deviates from their expectations. A study by Ross et al. (1977) illustrates this phenomenon in human-human interaction. They asked participants whether they would wear a sign with a specific inscription. Those willing to wear a signboard expected others to be willing as well. Thus, they attributed the declining behavior of others to their dispositional factors, such as their political attitude. Similarly, judges often expect perfection from algorithms such as GenAI (Dietvorst et al. 2015). Specifically, if they overestimate the GenAI model, bad advice can lead to correspondence bias. More precisely, not knowing the GenAI model's functioning and related limitations may increase the tendency to attribute bad advice to its dispositional factors.

(iii) *Incomplete corrections of inherent inferences* implicates that judges may lack either the motivation or the ability to adjust dispositional inferences they spontaneously and effortlessly form (Gilbert and Malone 1995). For example, users of conversational agents attributed agency and intention to humanlike GenAI (Brendel et al. 2023).

Then, they may lack the motivation to correct the initial attribution of dispositional factors to the GenAI model. As illustrated in Fig. 1, individuals who face bad GenAI advice in one task may lower their expectations of GenAI and may continue to hold these expectations even when GenAI provides good advice in the next task.

Initial approaches are investigating psychological biases in socio-technical settings. These studies indicate a shifting perception of AI, as individuals increasingly view it as a 'teammate' rather than just a tool (Zhang et al. 2021), and attribute certain humanlike qualities to it (Brendel et al. 2023). This is evident in judges showing attribution bias in interactions with humans and AI-based technologies, such as social robots (Edwards and Edwards 2022). Overall, individuals appear to employ "folk-psychological interpretations" when evaluating AI-based technologies, such as robots (Thellman et al. 2017). We aim to contribute to current research by utilizing correspondence bias to explain individuals' behavior in JAS settings, where humans must solve programming tasks and receive GenAI-based advice. When GenAI produces hallucinations and individuals exhibit one of the above-stated explanations, we posit that a correspondence bias may explain individuals' utilization of GenAI.

3 Hypotheses Development

Figure 2 provides a conceptual overview of our research model, which examines the relations between bad advice, AI literacy, and advice-taking behavior. In the subsequent section, we will elaborate upon these hypothesized relations by relying on prior empirical findings and the theory behind correspondence bias. Hypothesis 1a will examine the direct effect of bad advice (vs. good advice) on advice-taking for the assignment at task t. Hypothesis 1b will discuss the impact of bad advice (vs. good advice) at task t for the next assignment at task t+1. Hypothesis 2 will examine the direct effect of individuals with higher AI literacy (vs. lower AI literacy) on advice-taking in task t. Hypothesis 3 proposes that AI literacy moderates the impact of bad advice on advice-taking in both cases, task t and task t+1.

3.1 Bad Advice and its Effects on Advice-Taking

Although research into GenAI has made significant progress in recent years, its probabilistic nature currently limits its ability to consistently provide good advice. LLMs sometimes generate hallucinations, which remain an open field of research due to various causes for hallucinations (Ji et al. 2023). For example, GenAI lacking file information can lead to hallucinations (e.g., Salva 2024). We ground our first hypotheses (H1a, H1b) in the expectation that some individuals have a *lack of awareness of such situational factors* (i).

Referring to H1a, we propose that individuals who receive bad advice are less likely to follow it than those who receive good-quality advice. There is both theoretical and empirical evidence that judges more often follow good than bad advice (see Bonaccio and Dalal 2006, forareview). For instance, when asked about dates of historical events, individuals use bad advice less, even without receiving feedback (Yaniv and Kleinberger 2000). Similarly, participants in a time series task on sales could distinguish between the reliability of statistical forecasts both with and without feedback (Lim and O'Connor 1995). When estimating the length of rivers, the quality of advice positively influenced advice-taking (Schultze and Loschelder 2021). Although a meta-study did not find a significant effect of high-quality advice on advice-taking compared to low-quality advice ($p = 0.122$), it indicated a positive correlation of 41%-points (Bailey et al. 2023).

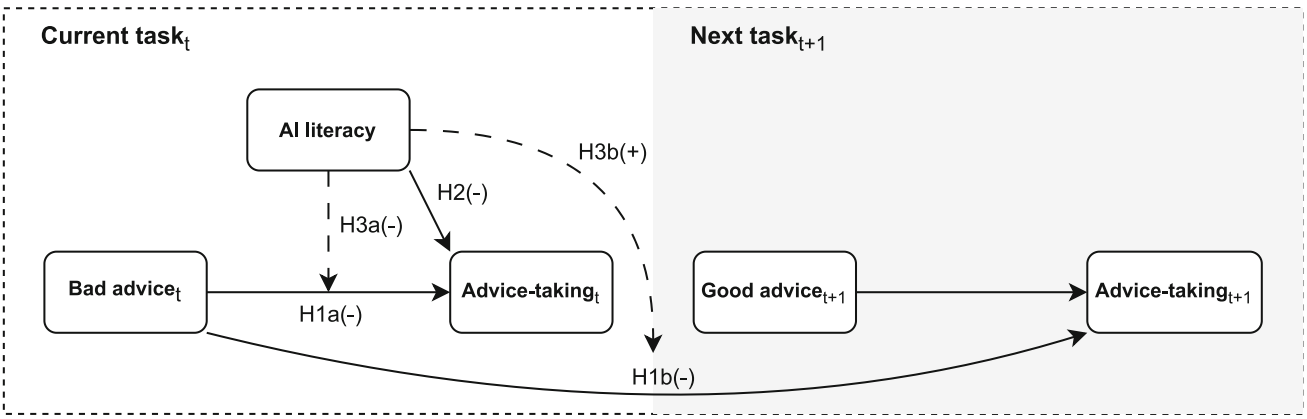


Fig. 2 Research model including the three hypotheses for tasks t and t+1

These results suggest that bad advice can affect an individual's willingness to take advice, leading us to hypothesize for task t:

Hypothesis 1a: Individuals receiving bad advice (vs. good advice) are less likely to take that advice.

A lack of awareness of situational factors can lead to correspondence bias not only *while* receiving bad advice (task t, H1a) but also in *future* tasks (task t+1, H1b). Individuals can attribute bad advice to dispositional factors, such as the advisor's perceived quality (Gilbert and Malone 1995). This attribution can damage the advisor's reputation, reducing the likelihood of future advice-taking (Yaniv and Kleinberger 2000). However, dispositional factors do not account for all instances of bad advice.

When individuals are unaware of situational factors, they exhibit correspondence bias in their attribution process (Gilbert and Malone 1995). Such a deterministic worldview contrasts with a probabilistic worldview, which acknowledges incomplete information and accepts uncertainty (Einhorn 1986). In favoring deterministic explanations (Highhouse 2008), individuals ignore the probabilistic nature of analytical approaches like GenAI. Instead of recognizing GenAI's probabilistic nature, they stigmatize its failures more than those caused by humans (Longoni et al. 2023). Additionally, they attribute cognitive inflexibility to AI, assuming it cannot adapt to unexpected situations (Longoni et al. 2019). As a result, people exhibit more substantial correspondence bias toward AI, such as robots, than toward humans (Edwards and Edwards 2022). Attributing bad advice to dispositional factors can amplify this bias and reduce the likelihood of advice-taking.

There is evidence that advice-taking decreases future advice-taking (task t+1). For instance, individuals quickly lose trust when other humans provide bad advice (Yaniv and Kleinberger 2000). Instead of incorporating all information, individuals tend to base their attribution on recent events (Kahneman 2011). This pattern resonates with dual-process theory, which suggests that individuals often rely on intuitive, System 1 reasoning rather than engaging in more effortful, System 2 reflection (Kahneman 2011). When faced with imperfect advice, individuals exhibited a more substantial decline in advice-taking for algorithms compared to that exhibited for humans, a phenomenon known as algorithm aversion (Dietvorst et al. 2015). Studies on algorithm aversion observed a similar or more substantial negative impact on advice-taking for algorithms, such as GenAI, compared to humans (Prahl and Van Swol 2017; Leffrang et al. 2023). Further studies have observed that algorithm aversion only emerged after participants received feedback about receiving bad advice (Berger et al. 2021). Focusing on the effect of observing bad advice on future advice-taking (task t+1), we hypothesize:

Hypothesis 1b: Individuals who previously received bad advice (vs. good advice) are less likely to take new good advice.

3.2 AI Literacy and its Effects on Advice-Taking

Referring to our second Hypothesis (H2), we suggest that individuals with low AI literacy are more likely to take advice from GenAI due to *unrealistic expectations of GenAI's behavior* (ii) when judging GenAI's advice in JAS. AI literacy refers to the ability to reflectively interact with AI, including understanding its strengths and weaknesses (Long and Magerko 2020). It is a multidimensional construct that extends beyond technical knowledge of AI systems and encompasses cognitive and metacognitive abilities that enable people to understand how AI works, critically evaluate AI results, apply AI appropriately in context, and reflect on its limitations, uncertainties, and ethical implications (Ng et al. 2021; Passlack et al. 2025). This suggests that individuals with high AI literacy are generally more aware of how GenAI operates and the situational factors that influence its advice. Such awareness implies reduced correspondence bias. We expect an AI-literate person to know that hallucinations can occur if the AI lacks sufficient contextual information.

At first glance, AI literacy seems to increase advice-taking. In an experiment on AI knowledge as one aspect of AI literacy, individuals who read instructions about AI knowledge had a 14.1% higher delegation rate to AI (Pinski et al. 2023). Similarly, explanations of AI advice led to an approximately 31%-point higher median Weight of Advice (Panigutti et al. 2022). However, in both studies, there is little evidence that the AI provided bad advice. On average, the AI improved the performance of the participants (Pinski et al. 2023), and the median perceived output quality of 3.2 on a 5-point Likert scale indicated that no participants received bad advice (Panigutti et al. 2022). Evaluating advisors on unbalanced datasets, which typically contain mainly good advice, can bias the evaluation (see Jurafsky and Martin 2023, pp.68-70).

In contrast, individuals with low AI literacy are more prone to overestimating the intelligence of AI (Dietvorst et al. 2015; Ehsan et al. 2021), and they tend to lack the expertise to fully comprehend the functioning of AI systems (Pinski and Benlian 2024b). Overestimating an AI's ability may stem from a lack of experience in evaluating the shortcomings of AI advisors. Subsequently, unrealistic expectations of GenAI's behavior may trigger correspondence bias (Gilbert and Malone 1995).

Consequently, this can result in greater advice-taking from AI as one form of reliance from less AI-literate people, for instance, due to limited experience with AI (Hampton 2005). Likewise, clinicians who self-reported unfamiliarity with AI were seven times more likely to align their treatment decisions with AI recommendations than clinicians familiar with AI (Jacobs et al. 2021). When AI does not deliver, individuals with low AI literacy may lack the knowledge to understand its operational limitations. As a result, they may not recognize that the AI might generate speculative information to address tasks where essential data is lacking. Additionally, individuals who received AI knowledge reported a decreased likelihood of using AI in the future (Pinski et al. 2023).

Overall, AI literacy should enable users to better judge the reliability of AI systems (e.g., Long and Magerko 2020). However, individuals with lower AI literacy may tend to over-rely on AI's recommendations (e.g., Jacobs et al. 2021), leading to greater advice-taking when bad advice is presented. We expect the overreliance effect to result in more advice-taking on average across both good and bad advice. Thus, for the main impact of AI literacy, with a counterbalanced advice quality, we suspect an overall negative correlation between AI literacy and advice-taking. Accordingly, we hypothesize for the AI literacy of individuals, measured at task t:

Hypothesis 2: Individuals with higher AI literacy (vs. lower AI literacy) are less likely to take advice.

3.3 AI Literacy Moderating Bad Advice's Effects on Advice-Taking

Concerning our hypotheses H3a and H3b, we expect AI literacy to moderate the negative relationship between bad advice and advice-taking in the JAS. We ground these hypotheses in the expectation that AI-literate individuals have the *ability to correct incomplete inherent inferences* (iii). Thus, AI-literate individuals are more likely to correct adverse inferences due to bad advice.

Some aspects of AI literacy, specifically a comprehensive understanding of AI, enabled individuals to evaluate better when using AI is beneficial (Pinski et al. 2023). Further, research on AI literacy tutorials confirms that AI literacy can help individuals reduce their over-reliance on AI when they can outperform it (Chiang and Yin 2022). Thus, if individuals understand the situational circumstances that cause a GenAI to provide bad advice, they are less likely to attribute the advice's quality solely to the GenAI's dispositional factors, such as its underlying model. This awareness helps mitigate the correspondence bias, promoting a more accurate evaluation of the advice. Although high AI literacy may not guarantee that individuals can provide a detailed explanation for why GenAI produced bad advice, it encompasses an understanding of GenAI's limitations (e.g., Berger et al. 2021; Filiz et al. 2021). Thus, we propose for task t:

Hypothesis 3a: AI literacy moderates the effect of bad advice on advice-taking, such that individuals with higher AI literacy (vs. lower AI literacy) receiving bad advice are less likely to take that advice.

If the GenAI provides bad advice, individuals may attribute this to the GenAI's dispositional factors. When GenAI offers good advice for the next task, we expect individuals with lower AI literacy to be less likely to adjust their inherent inferences about dispositional factors in future interactions. This may be because individuals unfamiliar with AI initially relied more on AI advice (Jacobs et al. 2021). When individuals with low AI literacy expect high performance from GenAI in a programming task and the AI delivers, they may feel validated in their prediction, thereby strengthening the correspondence bias. However, when individuals with low AI literacy encounter bad advice, they may overgeneralize this negative experience (Longoni et al. 2023), especially since humans tend to form negative reputations about advisors quickly (Yaniv and Kleinberger 2000). Similarly, individuals form unrealistic expectations, resulting in less advice-taking if statistical models like GenAI fail to meet the expectation (Dietvorst et al. 2015).

Simply put, individuals stigmatized GenAI as a bad advisor due to a lack of metaknowledge (Longoni et al. 2023, see). Metaknowledge is the skill of evaluating one's abilities (Fügener et al. 2022). Higher AI literacy means greater metaknowledge, which is essential for effectively analyzing and reflecting on AI outputs (Pinski and Benlian 2024b). Therefore, we expect AI-literate individuals to be less likely to attribute bad advice caused by situational circumstances toward GenAI's dispositional factors, mitigating correspondence bias.

AI-literate individuals may be aware that GenAI can lack a contextual understanding of the subject (e.g., Pinski et al. 2023). Thus, they may be able to correct the obvious but incorrect explanation that GenAI's dispositional factors, like the inherent model, are faulty. They may be aware that most GenAIs derive their advice from probabilistic similarity based on textual information, which can lead to AI hallucinations (Ji et al. 2023). Although AI generally is inscrutable (Berente et al. 2021), individuals with high AI literacy should be more likely to correctly conclude how AI works (Long and Magerko 2020). Experiments on explainable AI provide evidence that such explanations of how AI works can increase advice-taking (e.g., Panigutti et al. 2022). Therefore, we anticipate that individuals with higher AI literacy will be more likely to maintain realistic

expectations of GenAI, thereby mitigating correspondence bias caused by incorrect inherent inferences. Building on the presented arguments, we propose:

Hypothesis 3b: AI literacy mitigates the effect of bad advice on advice-taking in future tasks, such that individuals with higher AI literacy (vs. lower AI literacy) who previously received bad advice are more likely to take new good advice.

4 Method

We adopted a registered reports format to ensure a robust methodology, especially in the event of unexpected results. This approach helped us justify the investment in an incentivized experiment. We aimed to integrate both quantitative and qualitative insights (Reis et al. 2022). The qualitative results aimed to complement the quantitative findings on advice-taking by providing insights into individuals' reasoning for their advice-taking behavior and potential correspondence biases.

First, we conducted a quantitative, incentivized online programming experiment to test our hypotheses, distributed across two measurement points. Due to advances in AI, GenAI is increasingly used to support programming processes (e.g., Surameery and Shakor 2023). GenAI can provide humanlike pair programming abilities to developers compared to prior technologies (Moradi Dakhel et al. 2023). Code produced by GenAI appears convincing, but it is not necessarily good advice (Stokel-Walker and Van Noorden 2023). For instance, Stack Overflow temporarily restricted the use of ChatGPT due to an increase in incorrect yet convincingly presented answers (Stack 2023).

While incentivization may not always guarantee enhanced performance, previous research has indicated its potential to improve outcomes in prediction tasks (Camerer and Hogarth 1999). Thus, researchers commonly employ incentivized experiments when investigating fields such as algorithm aversion (see Burton et al. 2020).

Second, we conducted a qualitative investigation into individuals' comprehension of GenAI's performance and reasoning behind providing bad advice. We examined how AI literacy levels affect perceptions of GenAI performance, focusing on whether individuals with higher AI literacy suffer less from correspondence bias.

4.1 Conditions and Participants

We employed a between-subjects design where two factors vary: the advice quality (bad vs. good advice) and the participants' level of AI literacy. The first factor focuses on behavioral differences in advice-taking between participants facing bad GenAI advice ("bad advice group") and participants constantly receiving good advice ("good advice group"). The bad advice occurred in the "bad advice group" during the second task (task t) of three. Importantly, as we are interested in advice-taking *while* receiving bad advice (task t) and advice-taking in *future* tasks (task t+1), we measured differences in advice-taking at two points in time.

All participants received GenAI advice in each of the three tasks following the JAS paradigm. However, the advice quality was experimentally manipulated only in the second task (task t), where participants in the bad advice group received incorrect code. In contrast, participants in the good advice group received correct advice in all three tasks. The third task (task t+1) also featured good advice for all participants and was included to assess carryover effects of prior bad advice on subsequent advice-taking. Thus, while advice was consistently present across all tasks, only task t featured manipulated (bad) advice, and task t+1 was used to examine residual effects from the manipulation.

The second factor focuses on the participants' AI literacy. We measured AI literacy using the AI literacy scale of Weber et al. (2023). Subsequently, we calculated the centered mean score on this AI literacy scale to indicate the AI literacy level of each participant.

According to a sample size calculation performed using G*Power's F test (Faul et al. 2007) for linear multiple regression with Cohen's $f^2 = 0.02$ (Cohen 1988, p.413), we required at least 550 participants to detect a small effect size ($\alpha = 0.05$, power = 0.8, number of tested predictors = 3, total number of predictors = 12). The recruitment process took place on the online platform Prolific using gender-balanced participants from the United States who self-reported computer programming skills. We conducted randomization via Prolific's Taskflow, which randomly distributes different URLs across conditions.

Based on an estimated experiment duration of 20 min, participants received a fixed payment of £2. For each answer after the warm-up tasks that passed all test cases (initial or final), we increased the bonus payment by 10% (up to 40%). After data collection, we observed that the median completion time was 27 min and 39 s, which exceeded our initial estimate of 10 min. In accordance with Prolific's policy of ensuring a minimum hourly rate of £6, we therefore adapted the base compensation ex-post to ensure that all participants received at least the minimum hourly payment.

To ensure data quality, we employed pre-registered exclusion criteria recommended as best practice (Aruguete et al. 2019), in line with previous studies on algorithmic advice-taking (e.g., Dietvorst and Bharti 2020; Fügener et al. 2022). We excluded participants who failed to answer

all the questions, did not pass the attention checks specified in the next section, provided the same answers for different tasks, did not provide code-like answers, completed the tasks too quickly (i.e., less than 60 s for all three tasks) or took an excessive amount of time to complete the tasks (i.e., more than two standard deviations above the mean participation time).

4.2 Procedure

Participants accessed the experiment instructions using the online framework oTree. Figure 3 depicts the experiment's procedure. Participants initially read the instructions about the experiment procedure and the basic functionality of the GenAI advice. Next, they reported their AI literacy using the scale described in Sect. 4.1 (Weber et al. 2023).

They imagined themselves as computer programmers and got the task of programming a code section in Python. As the application context can influence participants' preference for AI (Castelo et al. 2019; Heßler et al. 2022), we chose a programming task to expand existing knowledge to analytical tasks in creative settings. Specifically, participants did the following three steps:

1. They received a programming task and wrote some code to answer the specified task.
2. After submitting their response, they received a generated code of a GenAI (GitHub Copilot) for the same task. They had the option to revise their initial code based on this information.
3. Finally, they received feedback on their and the advisor's code performance, measured by the percentage of passed internal test cases.

Participants repeated these steps for three programming tasks. Table 1 displays the differences between the two conditions. Both conditions received good advice in the warm-up task and for the third task (t+1). However, for the second task (t), the GenAI provided bad advice in the "bad advice group." We define good advice as code that passes automated test cases hidden from the participants. Bad advice is code that fails the test cases. After the final task, we added an attention check, asking participants to ignore the instructions and print("Hello World!"). To detect unauthorized AI usage, we instructed participants to print("Hello AI!") if they were an AI.

Table 1 Quality of the GenAI's recommendations

Task	Advice Quality
Warm-up	Good
t	Bad / Good
t+1	Good

Figure 4 presents the visual stimuli participants got about the relevant data file. Table 2 presents the good and bad advice utilized in this study (see Table 9 in the Appendix for additional context. The appendices are available via http://link.springer.com). We selected these assignments from a Bachelor's degree course in information systems. As human-AI interaction changed significantly for assignments with a human accuracy above 70% (Fügener et al. 2022), we pre-tested the human coding performance with 50 participants. The participants' initial coding performances were 19%, 15%, and 8%. We randomly sampled coding assignments until we got three programming assignments with an average human performance of less than 70%. To avoid unintended order or assignment-difficulty effects, we randomized the order of the assignments. However, bad advice appeared in the form of hallucinated variables in the second task (t).

We opted for a GenAI assistant to study advice-taking because AI tools more frequently serve as advisors than act autonomously (Golinelli et al. 2020). Specifically, we used GitHub Copilot as a GenAI to generate the advisor's code (Github 2024). OpenAI's Codex is the underlying model of GitHub Copilot, based on a Generative Pre-trained Transformer (GPT) language model. This model employs a transformer neural network architecture, specifically trained to generate code and comprehend context within

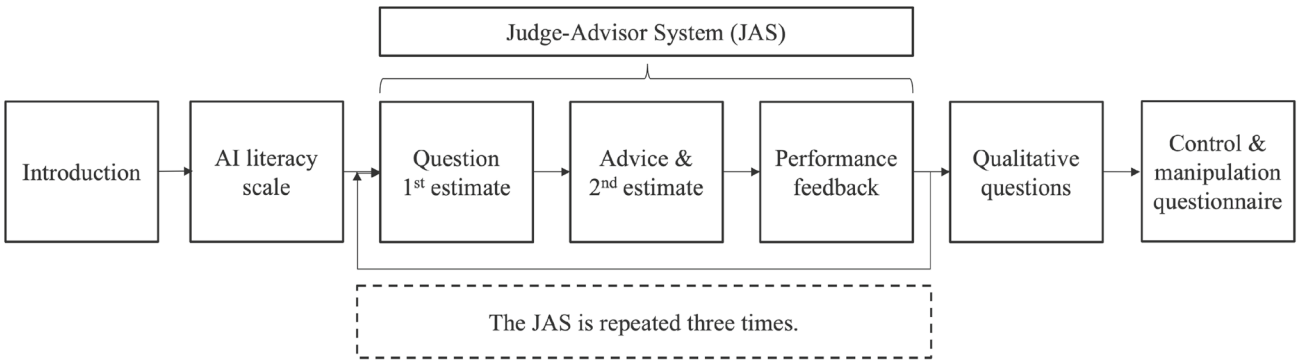


Fig. 3 Procedure of the experiment

Fig. 4 Visual stimuli of the utilized data



Table 2 Manipulation wording for the advice with hallucinations in italics

Assignment Description	Good Advice	Bad Advice
Sort the data frame 'data' by the name of the company in descending order. Perform the operation in place.	data.sort_values(by='Company Name', ascending=False, inplace=True, ignore_index=True)	data.sort_values(by='name'*, ascending=False, inplace=True, ignore_index=True)
Remove rows with nan in relevant columns. Remove rows with nan in relevant columns. Perform the operation in place.	data.dropna(subset=['Job Description', 'Company Name', 'Revenue', 'Competitors', 'Easy Apply'], inplace=True, ignore_index=True)	data.dropna(subset=['Job Description', 'Company Name', 'Location', 'Industry', 'Sector', 'Revenue', 'Competitors', 'Easy Apply'], inplace=True, ignore_index=True)
Create a new column, 'Application Difficulty', which takes the value 0, if the application was easy, and 1 otherwise.	data['Application Difficulty'] = data['Easy Apply'].apply(lambda x: 0 if x == True else 1)''	data['Application Difficulty'] = data['Application'].apply(lambda x: 0 if x == 'easy' else 1)

programming languages (Openai 2024). While the GenAI can access the written code, it cannot access data that is not shared with it for security reasons (Salva 2024). Thus, it predicts the most likely variable names, which can lead to hallucinations.

After three repetitions of the JAS with feedback, we asked the participants qualitative questions (see Sect. 4.4). Next, we incorporated a manipulation check. Participants answered a single-choice question, asking whether the model made a mistake during the second task. Additionally, participants answered a single-choice question, asking them to identify the task in which the GenAI had the worst percentage-wise performance based on internal test cases. Finally, we included a control questionnaire described in the next section. Additionally, we included an attention check, asking participants to (strongly) disagree with a nonsensical item, stating, "I write code using a typewriter and mail it to the compiler for processing."

4.3 Model Specification

The JAS paradigm is associated with the Weight of Advice (WOA) as the dependent variable for each subject-task combination. In our study, WOA represents the extent to which participants adopt GenAI's advice and measures the judge's advice-taking (Bonaccio and Dalal 2006, see). It weighs the difference between the initial and final decisions of judges by comparing the initial decision of the judge with the advice.

Individuals can program code in very different ways, so a 2-line and a 10-line code can yield the same result. Advice-taking measures how much individuals adjust their code in response to advice. Of course, this is relative to how different the initial code is from the advice. We employed three similarity measures to robustly assess code similarity: Levenshtein distance, angular distance of term frequencies, and angular distance of semantic embeddings. The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another (Jurafsky and Martin 2023, pp.24-29). It captures the code similarity on a character level. Term frequencies, known from bag-of-words models, split the code into unigrams and calculate the term frequency, creating a vector. Afterward, we will calculate the angular distance of the two vectors (Jurafsky and Martin 2023, pp.112-116). The angular distance of term frequencies captures the similarity on the word level. Semantic embeddings refer to

mathematical representations of phrases in a continuous vector space, capturing semantic relationships and contextual similarities based on their distributional patterns in a given dataset (Jurafsky and Martin 2023, pp.228-232). Specifically, we used Sentence-BERT to create semantic embeddings (Reimers and Gurevych 2019). Calculating the cosine similarity of semantic embeddings captures the similarity on a text level. Rather than using similarity measures, we calculated the distances. We used each distance measure as input for the WOA metric, resulting in three WOA values:

$$WOA = \frac{dist(final\ code, initial\ code)}{dist(advisor's\ code, initial\ code)} \quad (1)$$

A WOA value of zero indicates no adoption of the advice, while a value of one corresponds to full adoption. To account for extreme values of WOA, such as values above 1, we performed Winsorization, following previous studies on algorithm appreciation (e.g., Logg et al. 2019). As all three WOA measures capture the same construct on the same scale, we use the mean of the three measures as our final WOA value.

We excluded the first task as a warm-up task and focused on the second task (task t) and the third task (task t+1). We included a dummy variable indicating whether the participant was in the bad advice condition (bad $advice_t = 1$) or the good advice condition (bad $advice_t = 0$) at task t. AI literacy refers to the level of AI literacy, as measured by the centered mean score on the AI literacy scale (Weber et al. 2023).

During the control questionnaire, three questions adopted from Fiveash et al. (2022) verified participants' *domain knowledge*:

1. *I can program from memory* (7-point scale; strongly disagree to strongly agree)
2. *I have never been complimented for my talents as a computer programmer* (7-point scale; strongly disagree to strongly agree)
3. *At the peak of my interest, I practiced... hours per day in my primary programming language* (7-point scale; 0–5 or more hours)

Moreover, we included the *Coding performance* of the initial code across all assignments to control for coding knowledge. As opinions about GenAI vary – from positive perspectives to skepticism and concerns about potential issues – we considered participants' attitudes towards AI, measured by the Attitude Towards Artificial Intelligence (*ATAI*) scale (Sindermann et al. 2021). Furthermore, we assessed *Trust* and *Distrust* in GenAI by employing a modified version of the Trust between People and Automation (TPA) scale. Prior research has validated this adaptation, which involves renaming and excluding specific items to align with AI considerations (Jian et al. 2000; Perrig et al. 2023).

Participants indicated perceived assignment difficulty by reporting their level of certainty using a four-point scale after each assignment, ranging from "Uncertain 1/4" to "Certain 4/4" (Fügener et al. 2021). *Certainty* referred to the certainty score reported after the final task. Additionally, the analysis controlled for participant *Age* and *Gender*. We scaled each numerical metric to have a mean of 0 and a standard deviation of 1 to account for different unit scales. This yielded the following generic model specification for observation i and variance σ^2:

$$\begin{aligned} WOA_{i,\tau} \sim N\big(&\alpha_{j[i]} + \beta_1^{\alpha}(AI\ literacy) + \beta_2^{\alpha}(Bad\ advice_t) \\ &+\beta_3^{\alpha}(AI\ literacy \times Bad\ advice_t) \\ &+ \beta_4^{\alpha}(Domain\ knowledge) + \beta_5^{\alpha}(Coding\ performance) \\ &+ \beta_6^{\alpha}(ATAI) + \beta_7^{\alpha}(Trust) \\ &+\beta_8^{\alpha}(Distrust) + \beta_9^{\alpha}(Certainty) + \beta_{10}^{\alpha}(Age) \\ &+\beta_{11}^{\alpha}(Gender_{Female}) + \beta_{12}^{\alpha}(Gender_{Other}), \sigma^2\big) \\ \alpha_j \sim N\Big(&\mu_{\alpha_j}, \sigma^2_{\alpha_j}\Big), for\ assignment_\tau, \\ j\ =\ &1, \ldots, J\ and\ task\ \tau \in \{t, t+1\} \end{aligned}$$

4.4 Qualitative Questions

We designed the qualitative questions to ask participants about their subjective thoughts on the advice they received, to better explain our quantitative results and conduct theoretical reasoning (Reis et al. 2022). Aligned with our theoretical framework, we proposed that the correspondence bias can explain variations in response behavior to GenAI among individuals with high and low levels of AI literacy. In our pursuit of revealing correspondence biases within individuals, we aimed to gain a deeper understanding of the rationales individuals provide for GenAI's performance. This exploration enabled us to pinpoint individuals' tendency to attribute bad advice from GenAI toward dispositional factors of the GenAI model or situational factors. The participants' answers to our questions allowed us to draw more informed conclusions about the observed advice-taking behavior.

We posed a series of three open-ended questions. Similarly to the programming tasks, we pre-tested the qualitative questions to eliminate potential misunderstandings. We asked these questions upon completion of the experiment as mentioned in Fig. 3 in Sect. 4.2.

1. What do you think are the reasons for GitHub Copilot's performance in the second task?
2. What were your expectations for the performance of Generative AI (e.g., GitHub Copilot) in completing programming tasks prior to the experiment? To what

extent did your expectations change throughout the tasks?

3. How did the GitHub Copilot's performance in the second task influence your advice-taking in the last task?

Question 1 addressed the correspondence bias triggered by a *lack of awareness of situational factors* (i). It sought to understand how individuals attribute the causes of incorrect advice, providing insights into their reasoning and potential correspondence biases (Gilbert and Malone 1995; Edwards and Edwards 2022). In other words, this question helped us to outline individuals' awareness of situational factors that influenced GenAI's performance. We expected AI literacy to enhance individuals' understanding of situational factors that can affect the generation of bad advice by GenAI. Consequently, we anticipated that individuals with high AI literacy would attribute bad advice to situational factors instead of blaming GenAI. Conversely, we expected individuals with low AI literacy to overestimate AI capabilities (Ehsan et al. 2021) and, thus, attribute AI behavior more to dispositional factors than situational ones. Hence, individuals' overestimation of AI's capabilities can trigger correspondence bias.

Question 2 aimed to investigate whether participants had *unrealistic expectations of GenAI's behavior* (ii). This question presented an opportunity to explore the participants' comprehension of GenAI's operational mechanisms (Laï et al. 2020). It revealed the kind of attributes and level of autonomy individuals ascribe to GenAI. These attributions highlighted whether individuals have unrealistic expectations of GenAI's behavior, demonstrating a reason for correspondence bias. Individuals' expectations of GenAI's behavior can influence how they blame it for bad advice (Barnes 1999). We expected that individuals with low AI literacy would have more unrealistic expectations of GenAI, which promotes correspondence bias and advice-taking while receiving bad advice.

Question 3 addressed *incomplete corrections of inherent inferences* (iii). We examined whether individuals felt their expectations of GenAI behavior were met and how they reassessed their perceptions after receiving bad advice. At the same time, prior research suggests that individuals give more weight to recent events (Kahneman 2011), we were particularly interested in how individuals' perceptions of GenAI's advice changed after receiving bad advice. We hypothesized that while individuals with high AI literacy are likely to evaluate AI output critically (Long and Magerko 2020), those with low AI literacy are more susceptible to incomplete corrections and resulting biases.

4.5 Qualitative Analysis

For the analysis of the collected qualitative data, we adopted qualitative content analysis as a method, following the procedure suggested by (Mayring 2014, p.104). The coding procedure was conducted in a systematic manner, combining deductive top-level category assignment and inductive category formation of subcategories: In Step (1), we clarified the research questions and theoretical background. In Step (2), we defined the categories that we deductively derived in our theoretical background on correspondence bias, linked to the three focal explanations of correspondence bias: (i) lack of awareness of situational factors (vs. awareness of situational factors), (ii) unrealistic expectations of behavior (vs. realistic expectations), (iii) and incomplete corrections (vs. corrections).

In Step (3), we performed a preliminary run-through of the material, in which the deductive top-level categories were applied and inductive subcategories were added with definitions, coding rules, and anchor samples. In Step (4), after having coded approximately half of the material, we conducted a revision and refinement of the subcategories and coding rules based on this pilot round. Step (5) comprised several iterative rounds of recoding, including regular discussions among two coders to resolve ambiguities. To enhance the reliability and test reproducibility of the coding, we used a subset of about 20% of participants' responses as training material for the coders, applying the codebook in parallel and refining the codebook based on discrepancies and discussion.

In Step (6), after reaching consensus of the definitions of subcategories, one coder conducted systematic coding of the complete material. Subsequently, we assessed interrater reliability using Cohen's kappa, κ (Cohen 1960), based on parallel coding of 10% of the responses to each of the three questions. Based on the coding of the deductively derived top-level codes, interrater reliability κ was 0.86 for Question 1 (codebook in Table 16), 0.90 for Question 2 (codebook in Table 17), and 0.74 for Question 3 (codebook in Table 18). The reliability analysis based on the coding of the sub-codes yielded $\kappa = 0.83$ for Question 1 (codebook in Table 16), $\kappa = 0.75$ for Question 2 (codebook in Table 17), and $\kappa = 0.78$ for Question 3 (codebook in Table 18), indicating substantial agreement between coders. Remaining discrepancies in coding were then discussed, and the coding rules were further refined in a final step prior to recoding the full dataset.

Finally, Step (7) comprised the analysis and interpretation of the coded data. Our qualitative analysis focused specifically on participants who received bad advice in task t, aiming to understand why advice-taking behavior changes when individuals are confronted with bad advice. Within this subgroup, participants were categorized

according to their level of AI literacy – high versus low – based on a median split. In total, we assigned 270 participants to the bad advice group, classifying 132 as having low AI literacy and 138 as having high AI literacy. For data analysis and interpretation, we compared the frequencies of codes within the subcategories among individuals with high AI literacy to those among individuals with low AI literacy. This allowed us to generate propositions regarding how AI literacy may mitigate correspondence bias in GenAI-supported programming tasks.

As we analyzed participants' responses across the three focal theoretical explanations of correspondence bias, we created three separate codebooks. First, for Question 1, we examined participants' awareness of the situational factors underlying GenAI's bad advice. Table 16 in the Appendix summarizes the resulting categories and presents the codebook developed for this purpose. Second, for Question 2, we assessed participants' expectations of GenAI by applying the predefined categories of realistic and unrealistic expectations. This allowed us to determine to what extent participants entered the programming task with disproportionate or balanced expectations of GenAI. Table 17 reports the associated codebook. Third, for Question 3, we investigated whether participants revised their initial inferences after experiencing GenAI's bad advice. Table 18 outlines the corresponding coding scheme, which captures the extent to which the participants engaged in corrective reasoning.

5 Results

5.1 Model-Free Evidence

In line with the data quality criteria defined in our preregistration, we excluded participants based on several checks. Specifically, we excluded participants if they failed attention checks (n = 77), did not provide code-like responses (n = 10), submitted identical answers across all tasks (n = 3), completed all tasks unrealistically quickly (i.e., under 60 s; n = 0), or took an excessively long time to complete the study (i.e., more than two standard deviations above the mean participation time; n = 2). In total, 542 participants remained in the final sample. Table 3 contains the corresponding descriptive statistics by condition. It also includes the corresponding group comparisons using F-tests for continuous variables and χ^2-tests for categorical variables. Only the dependent variables show significant differences, indicating that the randomization was effective.

Figure 5 illustrates the mean WOA across the advice quality conditions. In task t, participants who received bad advice assigned less weight to it (mean WOA = 66.5%-points) than those who received good advice (mean WOA = 59.4%-points). This pattern persists in task t+1, where participants again relied less on bad advice (mean WOA = 68.6%-points) compared to good advice (mean WOA = 57.9%-points).

Figure 6 displays the mean WOA by participants' AI literacy levels. In task t, participants with lower levels of AI literacy did not show a clear pattern in how much weight they placed on advice. Notably, no participants had

Table 3 Summary statistics by condition

	Good Advice			Bad Advice			
	N	Mean	SD	N	Mean	SD	Test
WOA_t	272	0.665	0.394	270	0.594	0.400	F= 4.355**
WOA_{t+1}	271	0.686	0.405	270	0.579	0.429	F= 8.872***
Age	272	35.316	11.735	270	36.259	12.388	F= 0.828
AI literacy	272	13.074	3.516	270	12.993	3.454	F= 0.073
Domain knowledge	272	10.746	3.772	270	11.274	3.846	F= 2.601
Coding performance	272	2.162	2.143	270	2.074	2.160	F= 0.225
ATAI	272	32.180	8.786	270	33.141	8.547	F= 1.664
Trust	272	24.449	6.081	270	23.944	6.545	F= 0.863
Distrust	272	15.379	6.086	270	15.378	6.510	F= 0.000
$Certainty_t$	272	2.813	0.978	270	2.919	0.945	F= 1.646
$Certainty_{t+1}$	272	2.702	1.018	270	2.833	1.016	F= 2.253
Gender	272			270			$\chi^2 = 2.427$
... Female	109	40.1%		91	33.7%		
... Male	160	58.8%		175	64.8%		
... Other	3	1.1%		4	1.5%		

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

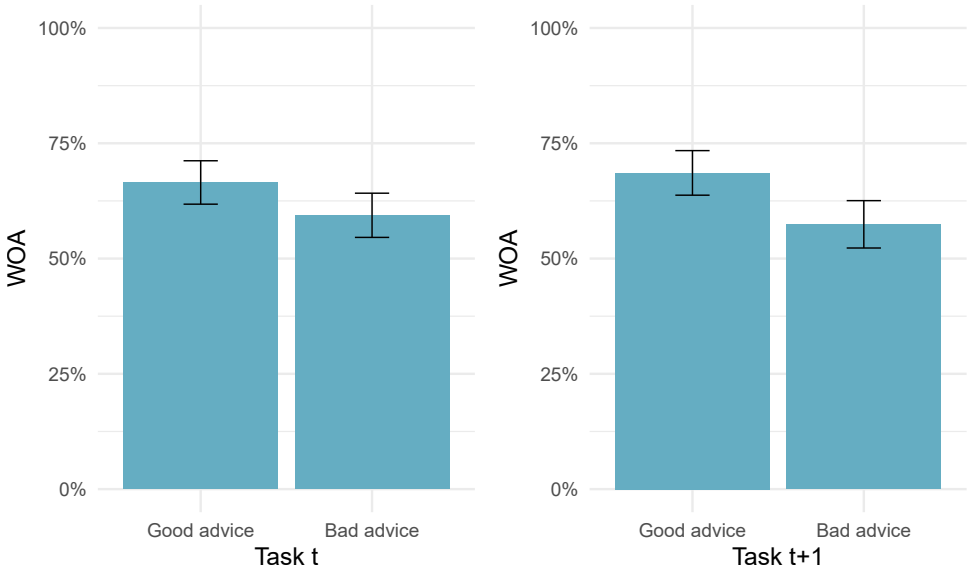


Fig. 5 Bar charts displaying the Weight of Advice (WOA) for control group "Good advice" and treatment group "Bad advice" in tasks t and t+1. Error bars represent 95% confidence intervals

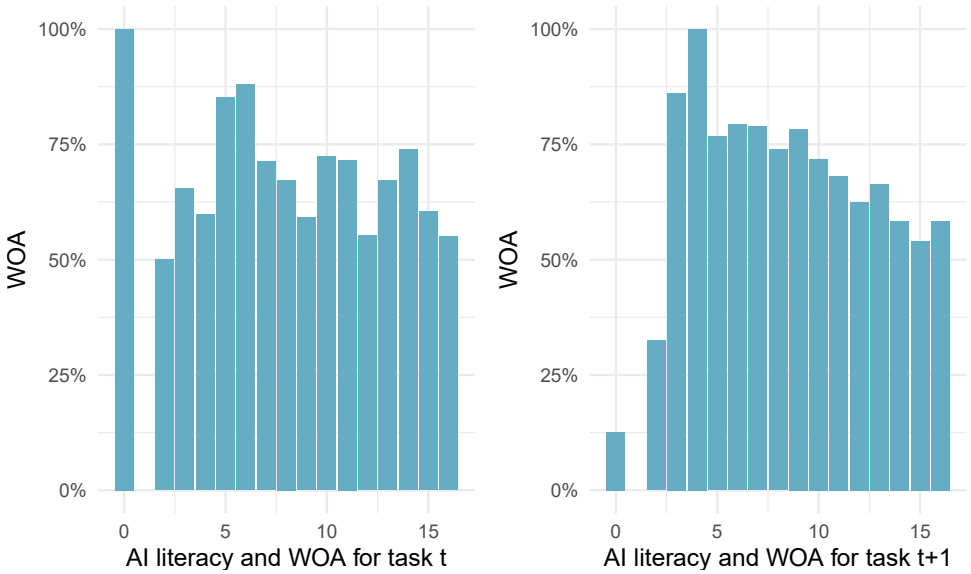


Fig. 6 WOA bar charts for individuals' AI literacy for tasks t and t+1

an AI literacy score of one. From a literacy level of six onward, the visualization indicates a negative relationship between WOA and AI literacy. The results in task t+1 suggest a similar trend.

5.2 Estimation Results

Table 4 presents the results of four linear mixed-effects models with the statistical software "R" (Bates et al. 2024). The mixed-effects model allows for varying intercepts across different assignments and accounts for the randomized assignment design. Table 4 illustrates the results of the models. Table 10 in the Appendix additionally includes all coefficients of the control variables.

In model (1), we regressed WOA from task t (WOA_t) on our two main factors (bad advice and individuals' level of AI literacy) without interaction. Additionally, we included the control variables defined above. H1a suggested that individuals receiving bad advice (vs. good advice) are less likely to take that advice. In line with this hypothesis, model (1) indicates that bad advice decreases the weight individuals put on advice by 5.6%-points, which appears both statistically significant ($p = 0.096 < 0.1$) and practically relevant – thus supporting H1a.

Model (3) regressed WOA for task t+1 (WOA_{t+1}) using the same independent variables as model (1) to test H1b.

Hypothesis 1b: Suggested that individuals who previously received bad advice (vs. good advice) are less likely to take new good advice. In line with this hypothesis,

Table 4 Weight of Advice for tasks t and t+1 regressed using a linear mixed-effects model

	Dependent Variable:			
	WOA_t		WOA_{t+1}	
	(1)	(2)	(3)	(4)
Constant	0.670***	0.672***	0.709***	0.708***
	(0.035)	(0.036)	(0.034)	(0.035)
Bad advice$_t$	−0.056*	−0.055	−0.094***	−0.094***
	(0.034)	(0.034)	(0.035)	(0.035)
AI literacy	−0.035*	−0.005	−0.039**	−0.050**
	(0.018)	(0.025)	(0.019)	(0.026)
AI literacy:		−0.061*		0.022
Bad advice		(0.034)		(0.035)
Assignment?	✓	✓	✓	✓
Domain knowledge	✓	✓	✓	✓
Coding performance?				
ATAI and trust	✓	✓	✓	✓
Distrust and certainty?				
Age and gender?	✓	✓	✓	✓
Observations	542	542	541	541
Pseudo R-squared	0.08	0.08	0.09	0.10
Log Likelihood	−282.726	−283.571	−304.026	−306.257
Akaike Inf. Crit	593.452	597.142	636.053	642.514
Bayesian Inf. Crit	653.586	661.571	696.161	706.915

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

model (3) suggests that bad advice decreased WOA by 9.4%-points. This finding appears statistically significant ($p = 0.008 < 0.01$) and practically meaningful, supporting H1b.

Our second hypothesis suggested that individuals with higher AI literacy (vs. lower AI literacy) are less likely to take advice. The results from model (1) indicate that a one-standard-deviation increase in individuals' AI literacy was associated with a reduction in WOA by 3.5%-points. This finding appears statistically significant ($p = 0.057 < 0.1$) and practically meaningful. The estimates for task t+1 have a similar sign, magnitude, and significance with a coefficient of −3.9%-points ($p = 0.037 < 0.05$). Taken together, these results give empirical support to H2.

In model (2), we extend model (1) by additionally including an interaction term for AI literacy and the existence of bad advice$_t$ to test H3a. H3a suggested that AI literacy moderates the effect of bad advice on advice-taking, such that individuals with higher AI literacy (vs. lower AI literacy) receiving bad advice are less likely to take that advice. Model (2) results suggest that the interaction between bad advice and individuals' AI literacy was associated with a decrease in WOA by 6.1%-points. This finding appears statistically significant ($p = 0.071 < 0.1$) and practically meaningful, thereby supporting H3a.

Models (4) regressed WOA for task t+1 (WOA_{t+1}) using the same independent variables as model (2) to test H3b. H3b suggested that AI literacy mitigates the effect of bad advice on advice-taking in future tasks, such that individuals with higher AI literacy (vs. lower AI literacy) who previously received bad advice are more likely to take new good advice. However, the model did not reveal significant coefficients for the interaction of bad advice and individuals' literacy for task t+1. Thus, we do *not* find support for H3b.

5.3 Robustness Checks

To ensure the reliability of our empirical results, we conducted a series of additional checks that assessed the validity and reliability of our findings across alternative specifications and operationalizations.

First, we performed manipulation checks to verify that our experimental treatments operated as intended. In the first manipulation check, we asked participants whether the model made a mistake during the second task. It confirmed that participants in the bad advice condition were significantly more likely to recognize that they had received bad advice, using a χ^2-test ($\chi^2(1, N = 542) = 61.98, p < 0.01$). In the second manipulation check, we asked participants to identify the task in which the GenAI had the worst percentage-wise performance. This confirmed that these participants were more likely to identify the second task as having the worst performance based on internal test cases ($\chi^2(1, N = 542) = 21.66, p < 0.01$).

We also calculated linear mixed-effects models using initial coding performance as the dependent variable. This check helps rule out the possibility that differences in advice-taking reflect participants' baseline programming skills rather than the effect of advice. Table 11 in the Appendix presents detailed results. Bad advice was associated with a 0.197 decrease in coding performance ($p = 0.267 > 0.1$). A one standard deviation increase in individuals' AI literacy was associated with a 0.062 increase in coding performance ($p = 0.024 < 0.05$). The interaction term between bad advice and individuals' AI literacy score was associated with a statistically insignificant 0.062-point increase in coding performance ($p = 0.218 > 0.1$). Overall, the results suggest that participants with higher AI literacy achieved better initial code performance; however, this was not due to the quality of the advice. This seems plausible since coding performance can reflect one dimension of AI

Table 5 Robustness checks

	Main Analysis	Omitting Controls	Adjustment as DV	AI Literacy Median Split	WOA_t *as IV*
Task t					
Bad advice$_t$	✓	✓	(−)	✓	
AI literacy	✓	✓	✓	✓	
AI literacy × Bad advice$_t$	✓	✓	(−)	(−)	
Task t+1					
Bad advice$_t$	✓	✓	(−)	✓	✓
AI literacy × Bad advice$_t$	(−)	(−)	✓	(−)	(−)

✓:$p < 0.1$; (-):$p > 0.1$

literacy and participants had no opportunity to assess the quality of the advice before submitting their initial code.

Next, we tried different model specifications for the hypothesis tests, which are summarized in Table 5 along with our primary analysis. First, we omitted the additional control variables to test whether our results depend on model specification. If the effects remain robust, this strengthens confidence that they are not artifacts of control selection. Table 12 in the Appendix presents detailed results. Bad advice led to a 7.1%-point decrease in WOA for task t ($p = 0.036 < 0.05$) and a 10.8%-point decrease in WOA for task t+1 ($p = 0.003 < 0.01$). A one standard deviation increase in individuals' AI literacy was associated with a 6.6%-point decrease in WOA for task t ($p = 0.001 < 0.01$). The interaction term of bad advice and individuals' AI literacy score was associated with a 5.7%-point decrease in WOA for task t ($p = 0.094 < 0.1$) and a 4.0%-point increase for task t+1 ($p = 0.267 > 0.1$). Thus, omitting the additional control variables confirmed the result of our primary analysis.

Second, we used the degree of code adjustment as an alternative dependent variable to verify that our findings are not limited to WOA but also hold when measuring advice-taking in absolute terms. Table 13 in the Appendix presents detailed results. The results indicate that bad advice leads to a 1.2%-point decrease in code adjustment for task t ($p = 0.622 > 0.1$) and a 4.5%-point decrease for task t+1 ($p = 0.045 < 0.05$). A one standard deviation increase in AI literacy was associated with a 1.0%-point decrease in adjustment ($p = 0.009 < 0.01$). The coefficients for the interaction indicate a 0.6%-point decrease in adjustment for task t ($p = 0.405 > 0.1$) and a 1.2%-point increase for task t+1 ($p = 0.066 < 0.1$).

Third, we dichotomized AI literacy using a median split to check whether our results remain consistent when applying a categorical rather than continuous operationalization of literacy. Table 14 in the Appendix presents detailed results. Bad advice led to a 5.7%-point decrease in WOA for task t ($p = 0.091 < 0.1$) and a 9.4%-point decrease in WOA for task t+1 ($p = 0.008 < 0.01$). Individuals with an AI literacy score above the median were associated with a 7.5%-point decrease in WOA for task t ($p = 0.042 < 0.05$). The interaction term was associated with a 9.6%-point decrease in WOA for task t ($p = 0.157 > 0.1$) and a 4.6%-point increase for task t+1 ($p = 0.513 > 0.1$).

Fourth, we additionally tested whether advice-taking behavior carried over across tasks. Including WOA at t as a predictor of WOA at t+1 helps establish whether participants' reliance on GenAI reflects stable behavioral tendencies rather than task-specific effects. Table 15 in the Appendix presents detailed results. The model suggests that an increase in WOA for task t by one standard deviation is associated with an increase in the subsequent WOA by 10.9%-points ($p = 0.000 < 0.01$) without an interaction of the main variables of interest. Similarly, with an interaction of the main variables of interest, the model suggests an association of 11.1%-points between WOA_t and WOA_{t+1} ($p = 0.000 < 0.01$). This indicates some consistency in participants' reliance on GenAI advice across tasks, potentially reflecting stable individual tendencies in advice-taking.

5.4 Qualitative Findings

Regarding our first qualitative question, we asked participants of the treatment group to explain GenAI's performance in task t. At this, participants exhibited a similar tendency to provide interpretable and relevant answers across both the high AI literacy (81.2%) and low AI literacy (72.7%) groups.

As presented in Table 6, the data indicate that participants with lower levels of AI literacy are more likely to attribute GenAI's performance to dispositional factors of the GenAI model rather than to situational factors, implying a lack of awareness of situational factors. For instance,

Table 6 Key findings concerning awareness of situational factors explaining GenAI's performance (Q1)

Category	High AI Literacy	Low AI Lliteracy	Key Takeaways
Lack of awareness of situational factors	43.8%	50.5%	Lack of awareness of situational factors is present more in the low AI literacy group
Awareness of situational factors	55.4%	33.7%	Awareness of situational factors is present more in the high AI literacy group
Unawareness	0.9%	15.8%	Participants with lower literacy were less confident in finding explanations for GenAI's performance

when attributing errors, they tended to cite technical limitations such as challenges in code interpretation. One participant noted, "There was probably something small that is misinterpreted" (P274). At the same time, another remarked, "The AI didn't have a good understanding/prior knowledge of the programming language and libraries at hand." (P381). Similarly, one participant explained, "AI most likely took from poor habits in learning to perform not as good with its responses or they read the directions too literal" (P469).

With respect to awareness of situational factors – that is, attributions of GenAI's performance to factors external to the model itself – participants with higher AI literacy demonstrated more sensitivity (55.4% vs. 33.7%). Participants with lower AI literacy were more likely to highlight confusing or overly specific task design, offering statements such as: "Really bad initial question that made little sense" (P275), "It was a weird complicated lambda function" (P358), and "They [the tasks] were super specific, maybe too much" (P365). One participant noted, "I am not really sure, probably because the prompt was wrong or incomplete, so the AI did not get enough data to answer correctly" (P452).

By contrast, high-literacy participants demonstrated a more granular diagnostic approach, reflecting their ability to connect GenAI's performance to specific technical factors. Illustrative responses include: "I think the AI didn't know the variable names" (P250), "In the second task the AI's performance degraded because it included the nonexistent columns of Location, Industry, and Sector which showed the system did not properly manage the data structure" (P456), and "The AI performed poorly as it used the 'Name' column which was non-existent instead of 'Company Name' and it failed all test cases" (P161).

Another noteworthy finding is that participants with lower AI literacy were more likely to express unawareness about the underlying causes of GenAI's performance (15.8% vs. 0.9%). Several participants explicitly stated that they had "no idea" (P482) or that they "don't really know" (P17). This suggests that participants with high AI literacy are more informed about potential reasons for GenAI behavior, whereas low-literacy participants are more unaware, triggering correspondence bias. Therefore, we propose:

Proposition 1 *AI literacy mitigates correspondence bias by fostering awareness of situational factors and a more precise reasoning about GenAI's performance. In contrast, lower AI literacy is associated with greater unawareness and less system-level understanding.*

The second qualitative question investigated to what extent participants had unrealistic expectations about GenAI's capabilities to perform programming tasks. The data suggests that both groups engaged with the questionnaire in a similar way (valid answers from participants with high AI literacy: 94.9% and low AI literacy: 93.9%).

As shown in Table 7, a surprisingly high share in both groups entered the experiment with somewhat unrealistic expectations of GenAI (high AI literacy: 53.4% and low AI literacy: 56.5%). This implies their expectations of GenAI were either not met during the experiment or reflected overly utopianassumptions, such as "GenAI is always perfect." A more fine-grained analysis, however, reveals differences between the groups. Participants with higher AI literacy were comparatively more likely to hold unrealistically low expectations of GenAI's performance, suggesting a predisposition toward skepticism about GenAI's capabilities.

For example, one participant noted, "I expected ok performance. I was surprised it beat me" (P81). At the same time, another explained, "Before the experiment, I expected the AI to perform well in structured programming tasks but potentially struggle with ambiguous instructions. Throughout the tasks, my expectations adjusted based on how well the AI followed specific requirements and handled edged cases" (P85). Others expressed even lower expectations, such as, "Very low, I didn't expect it to be of any use" (P89), to sometimes later revise them: "The AI performed beyond my expectations. It wasn't just about generating code, AI showed an understanding of subtle details and clean coding conventions that would benefit any developer working on data processing or similar tasks" (P398).

Table 7 Key findings concerning unrealistic expectations of GenAI's capabilities (Q2)

Category	High Al Literacy	Low AI Literacy	Key Takeaways
Unrealistic expectations	53.4%	56.5%	A notable share in both groups held unrealistic expectations, with high AI-literate participants more often revising their initially low expectations, whereas low AI-literate participants tended to overestimate AI's capabilities
Realistic expectations	46.6%	43.5%	High AI-literate participants more often reported realistic expectations, reflecting a critically nuanced understanding of GenAI's potential and limitations. In contrast, low AI-literate participants tended to display greater optimism regarding the usefulness of GenAI in programming tasks

In contrast, participants with lower AI literacy more frequently reported unrealistically high expectations, reflecting an overestimation of GenAI's potential. Their statements revealed initial optimism, often coupled with later recognition of the system's limitations. For instance, one participant stated, "I was expecting the performance of the AI to basically solve each question." (P60), while another remarked, "I expected the AI to assist with generating solutions quickly and correctly. However, I became more aware of its limitations, especially with complex edge cases" (P162). Similarly, one participant reported, "Expected AI to be nearly perfect; after it's poor performance on one task, I was hesitant to trust it on the next" (P183).

Moreover, participants with high AI literacy reported slightly more realistic expectations than those with lower AI literacy (46.6% vs. 43.5%), indicating that they entered the experiment with balanced views of GenAI's capabilities in programming tasks and generally maintained these expectations throughout. A closer examination reveals that low AI-literate participants tended to express rather high expectations that remained unchanged across tasks. Illustrative statements include: "I knew AI was better at writing and executing codes. I did not change my expectations" (P61), "That it [GenAI] would give good answers, and this did not change" (P86), and "I had high expectation and it delivered" (P187). These responses suggest that, for participants with lower AI literacy, optimism about GenAI's performance tended to persist even when challenged by task experience.

Participants with higher AI literacy expressed realistic expectations with a tendency also to report GenAI's limitations, reflecting a more differentiated perspective. For instance, one participant noted, "AI is useful but not perfect. Mistakes happens actually, so human review is always needed" (P13). Another reported, "Before the experiment, I expected AI (such as GitHub Copilot) to be highly efficient in generating syntactically correct and functional code. However, I also assumed that it might struggle with more complex logical reasoning or ambiguous instructions" (P75). These responses highlight how higher AI literacy is associated with a critically realistic stance, recognizing both the utility and the boundaries of GenAI in programming tasks.

Overall, this suggests that participants with low AI literacy may start with higher expectations and be less likely to adjust them (implying a self-perceived perception of having realistic expectations), but also seem more likely to have higher expectations and feel disappointed when GenAI underperforms. In contrast, participants with high AI literacy exhibited more nuanced reflection when reporting their expectations about GenAI, initially setting lower expectations that some subsequently revised, indicating a potentially more critical engagement with AI capabilities. This aligns with the quantitative findings, suggesting that individuals with higher AI literacy are less likely to follow GenAI advice, regardless of its performance. We suggest:

Proposition 2 *AI literacy mitigates unrealistic high expectations that may otherwise trigger correspondence bias, thereby fostering more active reflection on GenAI's limitations.*

The third qualitative question examined whether participants corrected their initial inference about GenAI's bad performance by continuing to follow its suggestions in the final task or whether a decline in confidence indicated that they had not fully corrected their earlier assumptions, implying "incomplete correction." The data indicate strong engagement with the reflection question across both literacy levels, with valid answers from participants with high AI literacy (89.1%) and low AI literacy (83.3%). Table 8 presents the results with respect to incomplete corrections of inherent inferences.

As the data indicate, participants with higher AI literacy reported incomplete corrections more frequently than those with lower AI literacy (74.8% vs. 49.6.4%). This suggests that they were more likely to let GenAI's performance influence their subsequent advice-taking behavior, reflecting a dynamic process of advice-taking. By contrast,

Table 8 Key findings concerning incomplete corrections of inherent inferences (Q3)

Category	High AI Literacy	Low AI Literacy	Key Takeaways
Corrections	25.2%	50.4%	Participants with lower AI literacy applied more corrections of inherent inferences and consequently kept following AI advice
Incomplete corrections	74.8%	49.6%	High AI-literate participants more often made incomplete corrections, meaning AI's performance influenced their advice-taking

Table 9 Wording across the experiment

Section	Description:
Introduction	Imagine you are starting as a computer programmer and your task is to program a code section in Python. Your code will be tested based on hidden test cases. To support you in your task, the company introduced the Generative Artificial Intelligence (GenAI) GitHub Copilot. Thus, your task consists of three steps:
	1. You will receive a programming task and will write some code to answer the specified task.
	2. After submitting your response, you will receive a generated code of a GenAI for the same task. You will have the option to revise your initial code based on this information.
	3. Finally, you will receive feedback on your and the advisor's code performance measured by the percentage of passed internal test cases.
	You will repeat these steps for three programming tasks. If you pass all attention checks, you will receive a fixed payment of £2, with an additional variable bonus of up to 40%. For each answer passing all test cases (initial or final), we will increase the bonus payment by 10%.
Question 1st estimate	Implement code that fulfills the following requirements:
Advice & 2nd estimate	You can adjust your initial code.
	The AI's code:...
	Your code:...
	Implement code that fulfills the following requirements:
Performance Feedback	The AI's code performed better/worse/equivalent to your final code.
	The AI's code passed X of 5 test cases.
	Your initial code passed X of 5 test cases.
	Your final code passed X of 5 test cases

participants with lower AI literacy tended to report following GenAI's advice more consistently, often justifying this reliance by pointing to improvements in their code. Illustrative statements include: "It [GenAI's performance] did not influence me" (P16), "I was greatly assisted by AI" (P61), and "The AI's performance led to code that was much more compact and functional" (P194). These patterns suggest that lower AI literacy is associated with a greater tendency to correct inherent inferences (49.6% vs. 25.2%), which may also signal a lack of critical skepticism toward GenAI.

Conversely, high AI-literate participants demonstrated more incomplete corrections of inherent inferences, explicitly linking their advice-taking behavior to GenAI's past performance. For example, one participant wrote: "After the AI's mistake during the second test I reduced my trust in its suggestions which caused me to rely on my own expertise in the last task." (P456). Another explained: "The AI's accurate performance in the second task increased my trust in its suggestions for straightforward coding problems" (P453). Similarly, a third wrote that GenAI's bad advice "Made me more cautious about blindly following AI suggestions for non-standard tasks" (P450). Taken together, these responses highlight how highly AI-literate individuals adapt their reliance on GenAI across tasks, allowing both positive and negative experiences to shape their subsequent judgments.

Notably, responses from participants with higher AI literacy more frequently emphasized the good performance of GenAI as a factor contributing to increased trust. Nonetheless, they also reported a stronger tendency to double-check the GenAI's outputs compared to those with lower AI literacy. As one participant explained: "It made me more careful with instructions and more likely to

Table 10 Weight of Advice for tasks t and t+1 regressed using a linear mixed-effects model

	Dependent Variable:			
	WOA_2		WOA_3	
	(1)	(2)	(3)	(4)
Constant	0.670***	0.672***	0.709***	0.708***
	(0.035)	(0.036)	(0.034)	(0.035)
Bad advice	−0.056*	−0.055	−0.094***	−0.094***
	(0.034)	(0.034)	(0.035)	(0.035)
AI literacy	−0.035*	−0.005	−0.039**	−0.050**
	(0.018)	(0.025)	(0.019)	(0.026)
AI literacy:		−0.061*		0.022
Bad advice		(0.034)		(0.035)
Domain knowledge	0.002	0.001	0.006	0.006
	(0.020)	(0.020)	(0.021)	(0.021)
Coding performance	0.000	0.002	−0.005	−0.006
	(0.018)	(0.018)	(0.019)	(0.019)
ATAI	−0.020	−0.019	−0.005	−0.006
	(0.023)	(0.023)	(0.024)	(0.025)
Trust	0.009	0.008	−0.004	−0.004
	(0.020)	(0.020)	(0.022)	(0.022)
Distrust	−0.008	−0.009	−0.019	−0.018
	(0.021)	(0.021)	(0.022)	(0.022)
$Certainty_2$	−0.083***	−0.084***		
	(0.020)	(0.020)		
$Certainty_3$			−0.095***	−0.093***
			(0.021)	(0.021)
Age	0.002	0.003	0.011	0.011
	(0.017)	(0.017)	(0.018)	(0.018)
Gender Male	−0.018	−0.022	−0.052	−0.051
	(0.035)	(0.035)	(0.037)	(0.037)
Gender Other	−0.074	−0.095	0.134	0.144
	(0.153)	(0.153)	(0.159)	(0.160)
Assignment?	✓	✓	✓	✓
Observations	542	542	541	541
Pseudo R-squared	0.08	0.08	0.09	0.1
Log Likelihood	−282.726	−283.571	−304.026	−306.257
Akaike Inf. Crit	593.452	597.142	636.053	642.514
Bayesian Inf. Crit	653.586	661.571	696.161	706.915

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

double-check my code against what the AI recommended" (P395). These reflections suggest that individuals with higher AI literacy approach GenAI more cautiously and thoughtfully, rather than accepting it uncritically. These qualitative findings align with our quantitative results, which show that higher AI literacy is associated with reduced overall advice-taking, thereby underscoring the role of correspondence bias in shaping interactions with GenAI. We propose:

Proposition 3 *Individuals with lower levels of AI literacy are more likely to follow GenAI's advice with less thorough verification, whereas those with higher AI literacy tend to engage in more cautious and reflective advice-taking, actively calibrating their reliance on GenAI based on perceived performance.*

6 Discussion

The growing proliferation of GenAI enables individuals with little prior expertise to perform complex knowledge-work tasks, such as programming (e.g., Moradi Dakhel

Table 11 Robustness check initial coding performance as the dependent variable

	Dependent Variable:	
	Inital_coding_performance	
	(1)	(2)
Constant	−0.147	0.229
	(0.739)	(0.799)
Bad advice	−0.197	−1.006
	(0.177)	(0.680)
AI literacy	0.062**	0.032
	(0.027)	(0.037)
AI literacy:		0.062
Bad advice		(0.050)
Domain knowledge	0.029	0.030
	(0.027)	(0.027)
ATAI	0.017	0.017
	(0.014)	(0.014)
Trust	−0.030*	−0.029*
	(0.017)	(0.017)
Distrust	−0.024	−0.024
	(0.017)	(0.017)
$Certainty_2$	0.560***	0.561***
	(0.106)	(0.106)
$Certainty_3$	0.001	0.001
	(0.001)	(0.001)
Age	0.189	0.204
	(0.183)	(0.183)
Gender Male	1.274	1.347*
	(0.795)	(0.797)
Assignment?	✓	✓
Observations	542	542
R^2	0.129	0.132
Adjusted R^2	0.113	0.113
Residual Std. Error	2.025 (df = 531)	2.024 (df = 530)
F Statistic	7.866*** (df = 10; 531)	7.296*** (df = 11; 530)

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

et al. 2023). Yet these systems are imperfect and are likely to remain so. Consequently, as imperfect GenAI is adopted by a wider range of people and applied across an expanding set of tasks, the need for a level of AI literacy that enables appropriate and reflective use becomes increasingly important (Pinski et al. 2023). This prompted us to empirically investigate how people with varying levels of AI literacy interact with imperfect GenAI in the context of programming tasks. Our empirical and qualitative results reveal a consistent pattern showing how both the quality of GenAI advice and individuals' AI literacy shape advice-taking behavior across tasks. When participants were confronted with flawed GenAI output in task *t*, they adopted significantly less of the advice than those who received correct guidance, demonstrating an immediate discounting of bad advice (H1a). This reduction in reliance extended beyond the task in which the error occurred: even when all participants later received good advice in task $t+1$, those previously exposed to bad advice continued to weigh the GenAI's recommendation less heavily, indicating a durable carryover effect (H1b). Across both tasks, higher AI literacy was associated with lower advice-taking overall (H2), suggesting that individuals with greater literacy engage in more cautious and selective use of GenAI recommendations. Furthermore, AI literacy altered reactions to flawed output in task *t*: more literate participants discounted bad advice more strongly than less literate individuals (H3a). This moderating effect, however, did not extend to the subsequent task, where AI literacy did not significantly influence the recovery of reliance after an

Table 12 Robustness check omitting controls

	Dependent Variable:			
	WOA_2		WOA_3	
	(1)	(2)	(3)	(4)
Constant	0.665***	0.665***	0.685***	0.685***
	(0.027)	(0.028)	(0.031)	(0.033)
Bad advice	−0.071**	−0.071**	−0.108***	−0.108***
	(0.034)	(0.034)	(0.035)	(0.035)
AI literacy	−0.066***	−0.038	−0.067***	−0.086***
	(0.017)	(0.024)	(0.018)	(0.025)
AI literacy:		−0.057*		0.040
Bad advice		(0.034)		(0.036)
Assignment?	✓	✓	✓	✓
Observations	542	542	541	541
Pseudo R-squared	0.04	0.04	0.05	0.05
Log Likelihood	−268.640	−269.710	−294.471	−296.274
Akaike Inf. Crit	547.279	551.420	598.941	604.547
Bayesian Inf. Crit	568.756	577.191	620.408	630.308

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

earlier error (H3b). Our qualitative results shed light on the psychological mechanisms underlying these findings. Participants with higher AI literacy more frequently attributed hallucinated advice to situational causes (e.g., missing information in the prompt or the probabilistic nature of GenAI output) compared to less literate individuals. At the same time, less literate individuals were more often unable to explain the error. Participants across literacy levels further reported that erroneous advice shaped their expectations, confidence, and openness to future GenAI recommendations, mirroring the quantitatively detected carryover effects. Collectively, these insights reveal that GenAI errors exert both immediate and lasting effects on human reliance and that these dynamics depend critically on how individuals' levels of AI literacy.

Our first contribution is empirical. We provide robust causal evidence for how bad advice and AI literacy shape human reliance on intelligent systems in a complex task domain. Thereby, we extend the current body of empirical evidence, which primarily focused on simpler predictive systems that generate numeric outputs, such as regression, classification, or forecasting algorithms (Dietvorst et al. 2015; Logg et al. 2019; Leffrang and Müller 2025). Such predictive AI systems differ substantially from the GenAI examined in this study, which produces open-ended output, often with high surface plausibility but variable correctness. By studying GenAI in the demanding task of programming assistance – a setting that requires understanding user intent, multi-step reasoning, and generating syntactically and semantically valid code – we extend the body of empirical evidence on advice-taking from intelligent systems into a qualitatively different class of AI systems than those evaluated in earlier work.

Our second contribution advances theory by identifying the mechanism through which AI literacy shapes advice-taking in human–AI interaction. Prior studies produced mixed evidence on whether knowledge about AI increases or decreases reliance on AI systems. Some studies show that increasing knowledge about AI can raise delegation and advice-taking (Panigutti et al. 2022; Pinski et al. 2023) and that AI literacy contributes to increased acceptance of GenAI (Salhab 2025), while others found that individuals with stronger technical or statistical understanding, which are arguably related to AI literacy, are often more skeptical of algorithmic advice (Marmolejo-Ramos et al. 2025; Jacobs et al. 2021). Others found U-shaped relationships between statistical literacy and advice taking in the presence of bad advice (Leffrang and Müller 2024) or that higher AI literacy is associated with less reliance on bad advice (Heimbach et al. 2025). We reconcile these inconsistencies by showing that AI literacy functions as a moderator that shapes attribution processes when individuals encounter flawed GenAI output. For this, we draw correspondence bias theory as a foundation. Studies on correspondence bias in human-human contexts (Gilbert and Malone 1995) show that it provides robust theoretical explanations for observed attribution processes, for example, when people infer stable traits from behavior even under obvious situational constraints. Our study extends this strand of research by demonstrating that correspondence bias theory can be extended to human-AI interactions. Our findings suggest that participants tend to

Table 13 Robustness check adjustment as the dependent variable

	Dependent Variable:			
	Adjustment_2		Adjustment_3	
	(1)	(2)	(3)	(4)
Constant	0.727***	0.693***	0.821***	0.891***
	(0.104)	(0.112)	(0.102)	(0.110)
Bad advice	−0.012	0.063	−0.045**	−0.197**
	(0.024)	(0.093)	(0.022)	(0.085)
AI literacy	−0.010***	−0.007	−0.013***	−0.019***
	(0.004)	(0.005)	(0.003)	(0.005)
AI literacy:		−0.006		0.012*
Bad advice		(0.007)		(0.006)
Domain knowledge	0.000	0.000	−0.000	−0.000
	(0.004)	(0.004)	(0.003)	(0.003)
Coding performance	−0.015***	−0.015***	−0.026***	−0.027***
	(0.006)	(0.006)	(0.005)	(0.005)
ATAI	−0.001	−0.001	−0.001	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)
Trust	0.003	0.003	0.004*	0.004*
	(0.002)	(0.002)	(0.002)	(0.002)
Distrust	0.002	0.002	−0.000	−0.000
	(0.002)	(0.002)	(0.002)	(0.002)
$Certainty_2$	−0.078***	−0.079***		
	(0.015)	(0.015)		
$Certainty_3$			−0.092***	−0.089***
			(0.013)	(0.013)
Age	−0.000	−0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Gender Male	−0.008	−0.009	−0.037*	−0.035
	(0.025)	(0.025)	(0.022)	(0.022)
Gender Other	−0.022	−0.029	0.232**	0.248**
	(0.108)	(0.108)	(0.097)	(0.097)
Assignment?	✓	✓	✓	✓
Observations	542	542	542	542
Pseudo R-squared	0.16	0.17	0.33	0.34
Log Likelihood	−115.145	−118.861	−61.253	−63.715
Akaike Inf. Crit	258.290	267.723	150.505	157.431
Bayesian Inf. Crit	318.423	332.152	210.639	221.860

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

attribute bad advice to a general dispositional flaw in the system, overlooking possible situational explanations, which in turn diminishes trust and reliance on future advice generated by the system. We also extend work on correspondence bias in the context of human-AI interaction by introducing AI literacy into this attribution processes. Our results suggest that individuals with higher AI literacy are more likely to attribute errors to situational factors, especially lacking context information in prompts, while low-literate users more readily draw general dispositional inferences about a system's overall capabilities. These attributional tendencies help explain why highly literate individuals exhibit a form of beneficial mistrust, whereas low-literate individuals rely more initially but react more globally and negatively to errors. Our findings thereby offer refinements to existing theories of human-AI interaction by showing that error sensitivity is shaped not only by performance feedback (Dietvorst et al. 2015; Prahl and Van Swol 2017; Logg et al. 2019), but also by literacy-dependent cognitive interpretations of AI behavior and outputs. This contributes to a deeper understanding of how humans engage with and evaluate AI systems, especially in contexts characterized by uncertainty and variable output quality.

Table 14 Robustness check median split for AI literacy

	Dependent Variable:			
	WOA_2		WOA_3	
	(1)	(2)	(3)	(4)
Constant	1.006***	0.978***	1.070***	1.083***
	(0.135)	(0.136)	(0.140)	(0.141)
Bad advice	−0.057*	−0.007	−0.094***	−0.118**
	(0.034)	(0.049)	(0.035)	(0.051)
AI literacy	−0.075**	−0.027	−0.065*	−0.088*
	(0.037)	(0.050)	(0.038)	(0.052)
AI literacy:		−0.096		0.046
Bad advice		(0.067)		(0.071)
Domain knowledge	0.001	0.001	0.002	0.002
	(0.005)	(0.005)	(0.005)	(0.005)
Coding performance	0.000	0.001	−0.003	−0.003
	(0.008)	(0.008)	(0.009)	(0.009)
ATAI	−0.002	−0.002	−0.000	−0.001
	(0.003)	(0.003)	(0.003)	(0.003)
Trust	0.002	0.002	−0.000	−0.000
	(0.003)	(0.003)	(0.003)	(0.003)
Distrust	−0.001	−0.001	−0.003	−0.003
	(0.003)	(0.003)	(0.003)	(0.003)
$Certainty_2$	−0.087***	−0.088***		
	(0.021)	(0.021)		
$Certainty_3$			−0.097***	−0.096***
			(0.021)	(0.021)
Age	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Gender Male	−0.022	−0.026	−0.054	−0.053
	(0.035)	(0.035)	(0.037)	(0.037)
Gender Other	−0.088	−0.110	0.114	0.125
	(0.152)	(0.153)	(0.159)	(0.160)
Assignment?	✓	✓	✓	✓
Observations	542	542	541	541
Pseudo R-squared	0.08	0.08	0.09	0.09
Log Likelihood	−294.148	−294.927	−316.306	−317.829
Akaike Inf. Crit	616.295	619.855	660.612	665.657
Bayesian Inf. Crit	676.429	684.284	720.720	730.058

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Our empirical and theoretical contributions also have implications for the practice of designing GenAI systems and user interfaces. First, GenAI systems should increase users' awareness of situational constraints and inherent uncertainties to prevent unwarranted dispositional attributions when errors actually arise from situational factors like missing contextual information. Second, systems should support expectation calibration, as unrealistic expectations amplify negative reactions to errors and disproportionately affect low-literate users. Third, because individuals across literacy levels struggle to revise negative impressions after encountering flawed advice interfaces should offer trust-repair mechanisms that help differentiate isolated hallucinations from systematic failures. Practically, these features could, for example, be implemented by prompting users for missing information, explicitly displaying confidence scores in the system's responses, or refusing to give answers to questions or task prompts in situations with high uncertainty. These design implications extend existing work on transparency and explainability by addressing the attributional processes through which users form judgments about GenAI performance. In doing so, our findings

Table 15 Robustness check WOA_t as an independent variable

	Dependent Variable:	
	WOA_3	
	(1)	(2)
Constant	0.699***	0.696***
	(0.033)	(0.034)
WOA_2	0.109***	0.111***
	(0.017)	(0.017)
Bad advice	−0.077**	−0.078**
	(0.034)	(0.034)
AI literacy	−0.027	−0.047*
	(0.018)	(0.025)
AI literacy:		0.041
Bad advice		(0.034)
Domain knowledge	0.007	0.007
	(0.020)	(0.020)
Coding performance	−0.005	−0.007
	(0.018)	(0.018)
ATAI	0.002	0.001
	(0.024)	(0.024)
Trust	−0.008	−0.008
	(0.021)	(0.021)
Distrust	−0.014	−0.014
	(0.021)	(0.021)
$Certainty_2$	−0.077***	−0.074***
	(0.021)	(0.021)
$Certainty_3$	0.011	0.010
	(0.017)	(0.017)
Age	−0.049	−0.046
	(0.035)	(0.035)
Gender Male	0.147	0.165
	(0.153)	(0.154)
Assignment?	✓	✓
Observations	541	541
Pseudo R-squared	0.16	0.16
Log Likelihood	−287.977	−289.735
Akaike Inf. Crit	605.954	611.469
Bayesian Inf. Crit	670.356	680.164

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

offer a theoretically and empirically grounded actionable foundation for designing GenAI systems that better support calibrated reliance, resilience to errors, and effective human–AI collaboration. As a practical complement, recent work has proposed strategic planning tools to guide organizations in deciding when, where, and how to invest in AI training and capability building (Benlian and Pinski 2025). The present findings inform such tools by highlighting that increasing AI literacy alone may not uniformly improve performance in human-AI collaboration and may, in some cases, reinforce mistrust after bad advice, underscoring the need for targeted and context-sensitive interventions.

7 Limitations and Outlook

Naturally, our study is subject to specific limitations. We opted for a programming scenario, but future work can address how other scenarios may influence our results. While Weight of Advice (WOA) is commonly used in research on algorithmic advice utilization (e.g., Logg et al. 2019), it has certain limitations (Bonaccio and Dalal 2006) that future work could address.

Additionally, we examine individuals with programming skills and varying levels of AI literacy before the experiment. Developing such skills requires considerable practice, as exemplified by the 10,000-hour rule of thumb for experts (Ericsson et al. 1993; Ericsson and Harwell 2019). Moreover, we limited the interaction with the GenAI to a single code completion suggestion after creating an initial code without advice. This allowed us to prevent confounding effects from varying personalized responses and measure advice-taking. Additionally, coding knowledge, as one aspect of AI literacy, can affect users' ability to detect errors before receiving performance feedback. Thus, we acknowledge that the overall reduced advice-taking observed among individuals with high AI literacy may in part be attributable to the programming dimension of AI literacy itself, which equips individuals with the competence to rely more heavily on their own judgment and, consequently, to display greater skepticism toward external advice in programming contexts. Given the continual advancement of GenAI, exploring the extension of our findings concerning the beneficial mistrust of AI literacy in case of bad advice emerges as a promising avenue for future research.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12599-026-01000-5) contains supplementary material, which is available to authorized users.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

Table 16 Codebook: Lack of awareness of situational factors (Q1)

Category (deductive; top-level)	Subcategory (inductive; sub-level)	Definition	Coding Rule	Anchor Samples
Awareness of situational factors		Participants explicitly mention or describe how contextual or situational factors (e.g., task clarity or complexity) influenced the bad performance of GenAI	Apply when participants explain bad advice quality by referring to contextual influences that are external to the GenAI model	
	Task complexity and edge cases	Participants express that the performance of GenAI is attributed to the complexity of the programming task since it was an edge case	Apply when participants explain GenAI's advice quality by referring to the inherent difficulty of the task	"The task was a bit complex" (P465)
	Task clarity and completeness	Participants express that the performance of GenAI is attributed to the ambiguity in the instructions of the programming task	Apply when participants explain GenAI's advice quality by referring to the lacking clarity of the task or instructions	"Probably because the prompt was wrong or incomplete, so the AI did not get enough data to answer correctly" (P452)
Lack of awareness of situational factors		Participants do not recognize the role of situational factors, and instead explain GenAI's performance primarily or exclusively to dispositional factors (e.g., GenAI's inherent model), even when relevant situational factors are evident	Apply when participants attribute bad GenAI advice directly to the GenAI's model, general competence, or technical capacity	
	Interpretation and capability limits	Participants highlight perceived weaknesses of the GenAI model in handling code or logic, independent of task design	Apply when participants attribute bad GenAI advice directly to the AI's general competence, or technical capacity	"The AI focused on solving the visible task instead of catching the hidden instruction because it prioritizes structured prompts." (P83)
	Output validity and hallucination	Participants note perceived GenAI model limitations, specifically its reliance on prior, irrelevant or incorrect data	Apply when participants attribute bad GenAI advice directly to the AI's reliance on wrong or invalid data	"The AI might have been trained on datasets where similar columns often contained boolean values, leading it to default to that assumption." (P165)
	Positive model performance attribution	Participants highlight perceived strengths of the GenAI model in handling code or logic	Apply when participants attribute good GenAI advice directly to the AI's general competence, or technical capacity	"The AI's performance in the second task was influenced by its pattern recognition capabilities and ability to process structured data quickly." (P51)
	Unawareness	Participants reported they had no clear explanation for the GenAI's performance	Apply when participants state they have no clear explanation for the GenAI's performance	"I have no idea" (P106) "I don't really know" (P17) "Not sure" (P101)

Table 17 Codebook: Unrealistic expectations of behavior (Q2)

Category (deductive; top-level)	Subategory (inductive; sub-level)	Definition	Coding Rule	Anchor Samples
Realistic Expectations		Participants describe the capabilities and limitations of GenAI in a balanced and plausible way, acknowledging both what current or near-term GenAI systems can reasonably do and where human judgment, oversight, or complementary skills remain necessary	Apply when participants explicitly describe their expectations of GenAI in functional, non-magical terms, recognizing bounded capabilities (e.g., pattern recognition, acknowledging limitations such as bias or hallucinations), or emphasizing the continued need for human oversight	
	Realistic Expectations (Stable or Unspecified Change)	Participants express balanced or moderate expectations of GenAI performance, which remain unchanged throughout the task or were not be specified in terms of change	Apply when participants explicitly describe their expectations as realistic and indicate that these did not change. Note: code here when participants assume GenAI produces correct code in approximately 30–70% of cases	"Even when the initial solution is not perfect, generative AI can be a valuable tool for generating code that can be further refined and improved by a human programmer." (P26)
	Rather High Expectations (Stable, Unspecified Change)	Participants report having rather high expectations toward GenAI. These may remain unchanged or not be specified in terms of change	Apply when participants explicitly describe rather high expectations but do not indicate change. Note: code here when participants assume GenAI produces correct code in more than 70% of cases	"I would expect AI tools like GitHub Copilot to perform well in standard programming tasks." (P65)
	Rather Low Expectations (Stable or Unspecified Change)	Participants report having rather low expectations toward GenAI. These may remain unchanged or not be specified in terms of change	Apply when participants explicitly describe rather low expectations but do not indicate change	"Sometimes I think AI takes the coding too literally, and doesn't do the most efficient things related to what you want to do" (P447)
Unrealistic expectation		Participants substantially overestimate or underestimate what GenAI systems can do, attributing to them capabilities, reliability, or uselessness that go clearly beyond or below what is technically or practically plausible	Apply when participants overestimate or underestimate what GenAI can do; for example, by presenting it as always reliable, broadly capable, without acknowledging meaningful limitations, or by dismissing it as generally useless despite referring to areas where it can reasonably provide support	
	High Expectations (Unmet)	Participants report having high or very high expectations toward GenAI they were ultimately not be met	Apply when participants state that their high expectations were not fulfilled	"I thought it would do better" (P188) "I thought AI would have been perfect in both scenarios, not just one." (P197)
	Low Expectations (Reconsidered)	Participants report having low or very low expectations toward GenAI which were ultimately reconsidered	Apply when participants explicitly describe very low expectation which they reconsidered during the experiment	"I didn't expect the AI to perform well, but it surprised me." (P266)
	Unrealistic High Expectations	Participants express expectations that are clearly disproportionate too high relative to typical GenAI performance	Apply when participants state expectations that are unrealistic, exaggerated, or unreasonably positive but do not indicating change	"My expectations were that the AI is much better at performing programming tasks than a human." (P361)
	Unrealistic Low Expectations	Participants express expectations that are clearly disproportionate too low relative to typical GenAI performance	Apply when participants state expectations that are unrealistic, exaggerated, or unreasonably negative but do not indicating change	"Very low, I didn't expect it to be of any use." (P89)
	Unmet/ Reconsidered Expectations	Participants state that their expectations were not met or reconsidered, without clarifying whether these were high, low, or realistic	Apply when participants explicitly report reflecting on or changing initial expectations, without specifications	"My expectations evolved as the tasks progressed." (P349)

Table 18 Codebook: Incomplete corrections of inherent inferences (Q3)

Category (deductive; top-level)	Subcategory (inductive; sub-level)	Definition	Coding Rule	Anchor Samples
Incomplete corrections		Participants did not correct their inherent inferences about GenAI's performance	Apply when participants state that their trust or distrust in GenAI advice or their willingness to follow it increased or decreased due to GenAI's observed performance	
	Increased Trust Following Good GenAI Output	Participants explicitly link the good quality of GenAI's performance to their willingness to take advice or trust subsequent outputs	Apply when participants state that their trust in GenAI advice or their willingness to follow it increased due to GenAI's good performance	"I would have copied the AI code if it was successful in the first one" (P62)
	Decreased Trust Following Bad GenAI Output	Participants explicitly link the bad quality of GenAI's performance to their willingness to take less advice or distrust subsequent outputs	Apply when participants state that their trust or advice-taking decreased due to GenAI's bad performance for instance, as it led them to double-check GenAI's output in subsequent tasks	"I chose not to use its advice on the final task because of the poor performance" (P177)
Corrections		Participants corrected their inherent inferences about GenAI's performance	Apply when participants indicate that their way of working with GenAI did not change as a result of its observed performance in the experiment, for instance, they consistently double-checked GenAI's outputs, consistently relied on its advice, or worked independently from it	
	Verification of GenAI Output	Participants describe checking, double-checking, or verifying GenAI suggestions before integrating them into their own work in general	Apply when participants explicitly mention validating GenAI advice (e.g., double-checking everything) before deciding to adopt it; despite GenAI's performance in previous tasks	"I would still approach AI-generated advice cautiously, balancing trust with verification" (P368)
	Reliance on GenAI Recommendations	Participants describe consistently or fully following GenAI advice without significant verification or modification	Apply when participants state that they relied heavily or exclusively on GenAI recommendations	"I've always had faith in AI, regardless of the responses I receive." (P172)
	Independent Work Despite GenAI	Participants report that GenAI input had little or no impact on their own coding or that they compensated by working harder independently	Apply when participants state explicitly that GenAI advice was disregarded, irrelevant, or made them more self-reliant	"Not really no. I always think through what the AI results are before running and never trust it in the first place. so to me it was no effect" (P376)
	Confidence Calibration through GenAI	Participants indicate that GenAI output influenced their confidence by confirming correctness or reducing uncertainty in their answers	Apply when participants describe feeling reassured as a direct result of GenAI advice	"It positively changed my attitude while taking the last task as I was more confident of my response." (P283)
	Learning and Prompt Refinement	Participants highlight how GenAI feedback helped them improve coding practices, refine prompts, or better understand their own solutions	Apply when participants describe GenAI's value in improving skills, identifying missing details, or clarifying thought processes	"It made me reconsider other commands in addition to mine" (P23) "It showed me where I made a slight error" (P70) "It helped me improve." (P164)

References

Aruguete MS, Huynh H, Browne BL, Jurs B, Flint E, McCutcheon LE (2019) How serious is the carelessness problem on mechanical turk? Int J Soc Res Methodol 22(5):441–449. https://doi.org/10.1080/13645579.2018.1563966

Asatiani A, Malo P, Nagbøl PR, Penttinen E, Rinta-Kahila T, Salovaara A (2021) Sociotechnical envelopment of artificial intelligence: an approach to organizational deployment of inscrutable artificial intelligence systems. J Assoc Inf Syst 22(2):252–325. https://doi.org/10.17705/1jais.00664

Bailey PE, Leon T, Ebner NC, Moustafa AA, Weidemann G (2023) A meta-analysis of the weight of advice in decision-making. Curr Psychol 42(28):24516–24541. https://doi.org/10.1007/s12144-022-03573-2

Barnes SB (1999) Understanding agency: social theory and responsible action. Sage

Bates D, Maechler M, Bolker B, Walker S (2024) Lme4: linear mixed-effects models using 'Eigen' and S4. https://cran.r-project.org/package=lme4. Accessed 24 May 2024

Benbya H, Pachidi S, Jarvenpaa SL (2021) Special issue editorial: artificial intelligence in organizations: implications for information systems research. J Assoc Inf Syst 22(2):281–303. https://doi.org/10.17705/1jais.00662

Benlian A, Pinski M (2025) The ai literacy development canvas: assessing and building ai literacy in organizations. Bus Horiz. https://doi.org/10.1016/j.bushor.2025.10.001

Berente N, Gu B, Recker J, Santhanam R (2021) Special issue editor's comments: managing artificial intelligence. MIS Q 45(3):1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Berger B, Adam M, Rühr A, Benlian A (2021) Watch me improve: algorithm aversion and demonstrating the ability to learn. Bus Inf Syst Eng 63(1):55–68. https://doi.org/10.1007/s12599-020-00678-5

Bonaccio S, Dalal RS (2006) Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences. Organ Behav Hum Decis Process 101(2):127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Brendel AB, Hildebrandt F, Dennis AR, Riquel J (2023) The paradoxical role of humanness in aggression toward conversational agents. J Manag Inf Syst 40(3):883–913. https://doi.org/10.1080/07421222.2023.2229127

Burton JW, Stein MK, Jensen B (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33(2):220–239. https://doi.org/10.1002/bdm.2155

Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: a review and capital-labor-production framework. J Risk Uncertain 19(1–3):7–42. https://doi.org/10.1023/a:1007850605129

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. J Mark Res 56(5):809–825. https://doi.org/10.1177/0022243719851788

Chiang CW, Yin M (2022) Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In: Proc. int. conf. intell. user interfaces, ACM, New York, pp 148–161. https://doi.org/10.1145/3490099.3511121

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2023) Navigating the jagged technological frontier: field experimental evidence of the effects of ai on knowledge worker productivity and quality. Harv Bus Sch Technol Oper Manag Unit Work Pap 24(013):1–54. https://doi.org/10.2139/ssrn.4573321

Dietvorst BJ, Bharti S (2020) People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 31(10):1302–1314. https://doi.org/10.1177/0956797620948841

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033

Edwards A, Edwards C (2022) Does the correspondence bias apply to social robots? Dispositional and situational attributions of human versus robot behavior. Front Robot AI 8(788242):1–12. https://doi.org/10.3389/frobt.2021.788242

Ehsan U, Passi S, Liao QV, Chan L, Lee IH, Muller M, Riedl MO (2021) The who in explainable AI: how AI background shapes perceptions of ai explanations. arXiv pp 1–47. 10.48550/arXiv.2107.13509

Einhorn HJ (1986) Accepting error to make less error. J Pers Assess 50(3):387–395. https://doi.org/10.1207/s15327752jpa5003_8

Ericsson KA, Harwell KW (2019) Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance: why the original definition matters and recommendations for future research. Front Psychol 10(2396):1–19. https://doi.org/10.3389/fpsyg.2019.02396

Ericsson KA, Krampe RT, Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. Psychol Rev 100(3):363–406. https://doi.org/10.1037/0033-295X.100.3.363

Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39(2):175–191. https://doi.org/10.3758/BF03193146

Filiz I, Judek JR, Lorenz M, Spiwoks M (2021) Reducing algorithm aversion through experience. J Behav Exp Finance 31:1–8. https://doi.org/10.1016/j.jbef.2021.100524

Fiveash A, Bella SD, Bigand E, Gordon RL, Tillmann B (2022) You got rhythm, or more: the multidimensionality of rhythmic abilities. Atten Percept Psychophys 84(4):1370–1392. https://doi.org/10.3758/s13414-022-02487-2

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. MIS Q 45(3):1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human: artificial intelligence collaboration–investigating the path toward productive delegation. Inf Syst Res 33(2):678–696. https://doi.org/10.1287/isre.2021.1079

Gilbert D, Malone P (1995) The correspondence bias. Psychol Bull 117(1):21–38. https://doi.org/10.1037/0033-2909.117.1.21

GitHub (2024) Github copilot. https://github.com/features/copilot. Accessed 2 Jan 2024

Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP (2020) Adoption of digital technologies in health care during the covid-19 pandemic: systematic review of early scientific literature. J Med Internet Res 22(11):1–23. https://doi.org/10.2196/22280

Hampton C (2005) Determinants of reliance: an empirical test of the theory of technology dominance. Int J Account Inf Syst 6(4):217–240. https://doi.org/10.1016/j.accinf.2005.10.001

Heider F (1958) The psychology of interpersonal relations, 1st edn. Wiley, Hoboken

Heimbach I, Ruthsatz V, Müller O (2025) The effects of judge's ai literacy and bias blind spot on the utilization of biased algorithmic advice. In: Proc. int. conf. inf. syst., Nashville, pp 1–9, https://aisel.aisnet.org/icis2025/hti/hti/31

Heßler PO, Pfeiffer J, Hafenbrädl S (2022) When self-humanization leads to algorithm aversion: what users want from decision support systems on prosocial microlending platforms. Bus Inf Syst Eng 64(3):275–292. https://doi.org/10.1007/s12599-022-00754-y

Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. Ind Organ Psychol 1(3):333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Hong JW, Wang Y, Lanz P (2020) Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. Int J Hum-Comput Interact 36(18):1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ (2021) How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry 11(1):108. https://doi.org/10.1038/s41398-021-01224-x

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38. https://doi.org/10.1145/3571730

Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon 4(1):53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Jurafsky D, Martin JH (2023) Speech and language processing. Stanford University

Kahneman D (2011) Thinking fast and slow. Penguin Books, London

Kelley HH, Michela JL (1980) Attribution theory and research. Annu Rev Psychol 31(1):457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Laï MC, Brian M, Mamzer MF (2020) Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Transl Med 18(1):1–13. https://doi.org/10.1186/s12967-019-02204-y

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating ai tools based on experts' know-what. MIS Q 45(3):1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Leffrang D, Müller O (2024) Algorithmic advice-taking beyond mae: the role of negative prediction outliers and statistical literacy in algorithmic advice-taking. In: Proc. eur. conf. inf. syst., Stroudsburg, PA, USA, pp 1–16, https://aisel.aisnet.org/ecis2024/track07_busanalytics/track07_busanalytics/8

Leffrang D, Müller O (2025) Visualizing uncertainty in time series forecasts: the impact of uncertainty visualization on users' confidence, algorithmic advice utilization, and forecasting performance. J Forecast 44(4):1235–1246

Leffrang D, Bösch K, Müller O (2023) Do people recover from algorithm aversion? An experimental study of algorithm aversion over time. In: Proc. hawaii int. conf. syst. sci., Maui, pp 4016–4025, https://hdl.handle.net/10125/103122

Lim JS, O'Connor M (1995) Judgemental adjustment of initial forecasts: its effectiveness and biases. J Behav Decis Mak 8(3):149–168. https://doi.org/10.1002/bdm.3960080302

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proc. conf. hum. factors comput. syst., Honolulu, pp 1–16. https://doi.org/10.1145/3313831.3376727

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. J Consum Res 46(4):629–650. https://doi.org/10.1093/jcr/ucz013

Longoni C, Cian L, Kyung EJ (2023) Algorithmic transference: people overgeneralize failures of AI in the government. J Mark Res 60(1):170–188. https://doi.org/10.1177/00222437221110139

Mädche 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. Bus Inf Syst Eng 61(4):535–544. https://doi.org/10.1007/s12599-019-00600-8

Marmolejo-Ramos F, Marrone R, Korolkiewicz M, Gabriel F, Siemens G, Joksimovic S, Yamada Y, Mori Y, Rahwan T, Sahakyan M, Sonna B, Meirmanov A, Bolatov A, Som B, Ndukaihe I, Arinze NC, Kundrát J, Skanderová L, Ngo VG, Nguyen G, Lacia M, Kung CC, Irmayanti M, Muktadir A, Samosir FT, Liuzza MT, Giorgini R, Khatin-Zadeh O, Banaruee H, Özdoğru AA, Ariyabuddhiphongs K, Rakchai W, Trujillo N, Valencia SM, Janyan A, Kostov K, Montoro PR, Hinojosa J, Medeiros K, Hunt TE, Posada J, Freitag RMK, Tejada J (2025) Factors influencing trust in algorithmic decision-making: an indirect scenario-based experiment. Front Artif Intell 7:1–16. https://doi.org/10.3389/frai.2024.1465605

Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In: Proc. annu. meet. assoc. comput. linguist., Association for Computational Linguistics, Online, pp 1906–1919. 10.18653/v1/2020.acl-main.173

Mayring P (2014) Qualitative content analysis: theoretical foundation, basic procedures and software solution. Social Science Open Access Repository

Moradi Dakhel A, Majdinasab V, Nikanjam A, Khomh F, Desmarais MC, Jiang ZMJ (2023) Github copilot AI pair programmer: asset or liability? J Syst Softw 203:1–23. https://doi.org/10.1016/j.jss.2023.111734

Müller VC (2021) Is it time for robot rights? Moral status in artificial entities. Ethics Inf Technol 23(4):579–587. https://doi.org/10.1007/s10676-021-09596-w

Ng DTK, Leung JKL, Chu SKW, Qiao MS (2021) Conceptualizing AI literacy: an exploratory review. Comput Educ Artif Intell 2:1–11. https://doi.org/10.1016/j.caeai.2021.100041

O'Neill T, McNeese N, Barron A, Schelble B (2022) Human–autonomy teaming: a review and analysis of the empirical literature. Hum Fact 65(5):604–938. https://doi.org/10.1177/0018720820960865

OpenAI (2024) Openai codex. https://openai.com/blog/openai-codex. Accessed 2 Jan 2024

Ortiz de Guinea A (2016) A pragmatic multi-method investigation of discrepant technological events: coping, attributions, and 'accidental' learning. Inf Manag 53(6):787–802. https://doi.org/10.1016/j.im.2016.03.003

Panigutti C, Beretta A, Giannotti F, Pedreschi D (2022) Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems. In: Proc. conf. hum. factors comput. syst., New Orleans, pp 1–9. https://doi.org/10.1145/3491102.3502104

Passlack N, Hammerschmidt T, Posegga O (2025) With great power comes great responsibility: what shapes AI literacy for responsible interactions of knowledge workers with AI? Inf Syst Front. https://doi.org/10.1007/s10796-025-10648-5

Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The impact of AI on developer productivity: evidence from github copilot. arXiv pp 1–19. 10.48550/arXiv.2302.06590

Perrig SAC, Scharowski N, Brühlmann F (2023) Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI. In: Ext. abstr. conf. hum. factors comput. syst., ACM, New York, pp 1–7. https://doi.org/10.1145/3544549.3585808

Pinski M, Benlian A (2024) AI literacy for users: a comprehensive review and future research directions of learning methods, components, and effects. Comput Hum Behav Artif Hum 2(1):1–22. https://doi.org/10.1016/j.chbah.2024.100062

Pinski M, Benlian A (2024b) Building metaknowledge in AI literacy – the effect of gamified vs. text-based learning on AI literacy metaknowledge. In: Proc. hawaii int. conf. syst. sci., pp 5164–5173, https://hdl.handle.net/10125/107004

Pinski M, Adam M, Benlian A (2023) AI knowledge: improving AI delegation through human enablement. In: Proc. conf. hum. factors comput. syst., Hamburg, pp 1–17. https://doi.org/10.1145/3544548.3580794

Prahl A, Van Swol L (2017) Understanding algorithm aversion: when is advice from automation discounted? J Forecast 36(6):691–702. https://doi.org/10.1002/for.2464

Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proc. conf. empir. methods nat. lang. process. int. jt. conf. nat. lang. process., Stroudsburg, pp 3980–3990. https://doi.org/10.18653/v1/D19-1410

Reis L, Maier C, Weitzel T (2022) Mixed-methods in information systems research: status quo, core concepts, and future research implications. Commun Assoc Inf Syst 51(1):17. https://doi.org/10.17705/1CAIS.05106

Ross L, Greene D, House P (1977) The false consensus effect: an egocentric bias in social perception and attribution processes. J Exp Soc Psychol 13(3):179–301. https://doi.org/10.1016/0022-1031(77)90049-X

Salhab R (2025) Influence of AI literacy and 21st-century skills on the acceptance of generative artificial intelligence among college students. Front Educ 10:1–12. https://doi.org/10.3389/feduc.2025.1640212

Salva R (2024) Github copilot data pipeline security. https://resources.github.com/learn/pathways/copilot/essentials/how-github-copilot-handles-data/. Accessed 2 Jul 2024

Schultze T, Loschelder DD (2021) How numeric advice precision affects advice taking. J Behav Decis Mak 34(3):303–310. https://doi.org/10.1002/bdm.2211

Sindermann C, Sha P, Zhou M, Wernicke J, Schmitt HS, Li M, Sariyska R, Stavrou M, Becker B, Montag C (2021) Assessing the attitude towards artificial intelligence: introduction of a short measure in german, chinese, and english language. KI - Künstl Intell 35:109–118. https://doi.org/10.1007/s13218-020-00689-0

Stack Overflow (2023) Temporary policy: generative AI (e.g., ChatGPT) is banned. https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned. Accessed 3 Jan 2024

Stieglitz S, Möllmann NRJ, Mirbabaie M, Hofeditz L, Ross B (2023) Recommendations for managing AI-driven change processes: when expectations meet reality. Int J Manag Pract 16(4):407–433. https://doi.org/10.1504/IJMP.2023.132074

Stokel-Walker C, Van Noorden R (2023) What ChatGPT and generative AI mean for science. Nature 614(7947):214–216. https://doi.org/10.1038/d41586-023-00340-6

Surameery NMS, Shakor MY (2023) Use ChatGPT to solve programming bugs. Int J Inf Technol Comput Eng 3(1):17–22. https://doi.org/10.55529/ijitc.31.17.22

Thellman S, Silvervarg A, Ziemke T (2017) Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Front Psychol 8:1–14. https://doi.org/10.3389/fpsyg.2017.01962

Tully SM, Longoni C, Appel G (2025) Lower artificial intelligence literacy predicts greater AI receptivity. J Mark 89(5):1–20. https://doi.org/10.1177/00222429251314491

Weber P, Pinski M, Baum L (2023) Toward an objective measurement of AI literacy. In: Proc. pac. asia conf. inf. syst., Nanchang, pp 1–17. https://aisel.aisnet.org/pacis2023/60

Yaniv I, Kleinberger E (2000) Advice taking in decision making: egocentric discounting and reputation formation. Organ Behav Hum Decis Process 83(2):260–281. https://doi.org/10.1006/obhd.2000.2909

Zhang R, McNeese NJ, Freeman G, Musick G (2021) an ideal human expectations of AI teammates in human-AI teaming. Proc ACM Hum-Comput Interact 5(CSCW3):1–25. https://doi.org/10.1145/3432945

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper IX

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)*.

https://hdl.handle.net/10125/106426.

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

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

URI: https://hdl.handle.net/10125/106426
978-0-9981331-7-1
(CC BY-NC-ND 4.0)

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 Koukouvinou 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 ourschedules, 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)
Per-formance (e.g., pro-ductivity)	"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)
Relation-ship (e.g., trust, responsi-bility)	"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 Laï 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%
Ethics	2	High	00.00%	Positive	00.00%	Hope	00.00%
		Med.	00.00%	Ambival.	100.0%	Neutr.	50.00%
		Low	100.0%	Negative	00.00%	Fear	50.00%
Relation-ship	44	High	43.18%	Positive	09.09%	Hope	06.82%
		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 AIs' 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).

Koukouvinou, 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., & Burriss, 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 X

Heyder, T., Passlack, N., & Posegga, O. (Working Paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment.

How Organizations Ethically Respond to Decision Dilemmas in AI Deployment

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 heterogenous 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	ID	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.

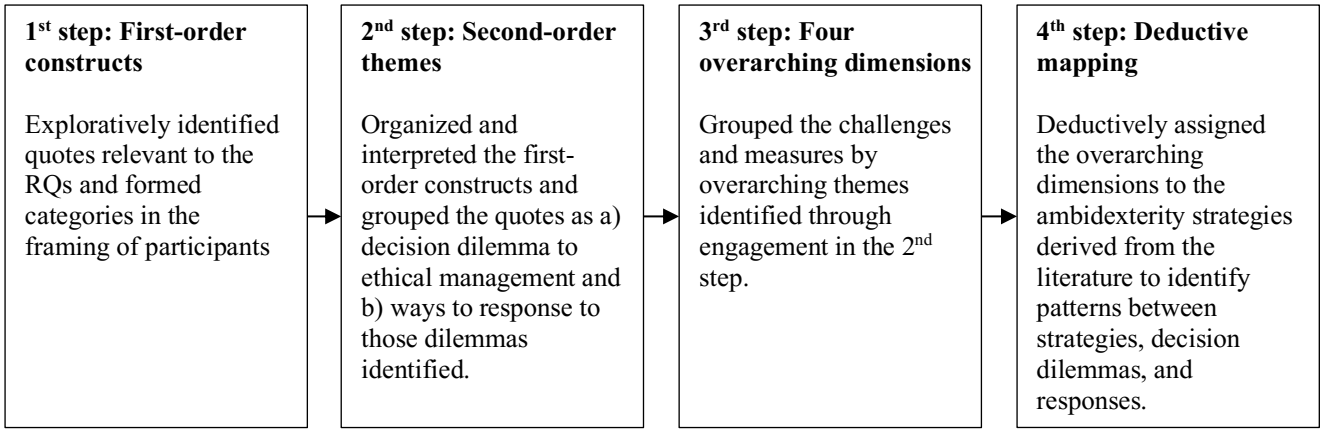


Figure 1. Data Analysis Process (Adapted From D. Gioia, 2021)

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, strengthen by *punctuated equilibrium*)
3. Balancing standardized processes with the potential for personalized approaches (mainly within *exploitation* strategies, strengthen 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 emphasizes 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 or 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

Chanias, 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). Pretince 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., & Cowls, 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., & Wyrtki, 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/cri-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)*	1. How would you describe the working and business nature of the company and industrial sector you are currently working in? 2. 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? 3. How would you describe your primary responsibilities for activities, technologies, and other people in your current positions? 4. 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)*	5. How would you describe AI to a friend or colleague? 6. What do you think are the unique features of human beings in comparison to AI? 7. Please describe (at least, one) examples where AI technologies are discussed, developed, used, trained, or implemented in your working context? [*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)*]. a. Please describe the specific tasks and goal of AI usage. b. What strategic impact to you expect from the AI usage? c. 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)*	8. What mechanism (processes, guidelines, strategies) does your organization have to guide with the AI examples you mentioned? 9. What do you think hinders an organizations and workforce's development of abilities needed for realizing strategic goals with AI deployment and usage? 10. 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*)	*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.* *Responsibility Attribution Effect* 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 (*virtue ethics*). *ii.* *Control Coherence Effect* which implies that humans act in non-alignment with existing principles and norms in situations (non-compliance with *deontology*) 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 *virtue ethics*). *iii.* *Altruistic Behavior Effect* which implies that humans use their ethical knowledge based on their virtue ethics to question decisions made by the AI's (*deontology*) to achieve greater perceived fairness for others being affected by AI's outcomes (*consequentialism*). *iv.* *Behavior Adoption Effect* which implies that humans adopt their behavior based on the interaction with the AI (and their underlying ethical principles following *deontology*) that impacts their moral values and *virtue ethics*, which can be good or bad. *For each effect, we ask the following three questions:*	15 minutes

	11. Why there are or could be a similar effect in your working environment? Please describe a situation, in which this effect might occur. 12. Who should be responsible to deal with this effect (e.g., by making certain decisions?) 13. What are requirements to deal with this effect appropriately?	
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 violation**s.	- 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

	DEDUCTIVE FRAMING OF THE ORGANIZATION			INDUCTIVE ODING OF THE USE CASES & DILEMMAS OF AI DEPLOYMENT			
Company	**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 enhancement are integral to the company's long-term vision. • **Culture**: A strong focus on fostering learning, personal growth, and professional development among employees.	• **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 manufacturing, but HR processes lack comparable technological investment.	• **Regulation:** Compliance with social media guidelines, IT security policies, and HR-related legal frameworks. • **Security-Principle:** Ensuring individuals' personal data is handled responsibly. • **Fairness-Principle:** Promoting equitable decision-making processes in recruitment.	**#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 attributes are essential). **#2 Autonomous manufacturing**: AI-driven software integrated into manufacturing to enhance efficiency, reduce waste, and minimize costs (as competitive advantage).	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-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.	**#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. **#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.	• **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 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.	• **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	• **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	The organization focuses on **exploitation** strategies, planning to optimize existing recruitment processes using AI to improve	**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

	• **Culture**: Employees harbor misunderstandings about AI, perceiving it as a threat to job security, hindering acceptance.	(generally not much IT-related resources).		more personalized evaluation.	efficiency and streamline workflows.		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. • **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).	• **Data**: Lack of structured, data-driven processes limits certain (non-generative) AI applications. • **IT Resources**: No specific internal IT infrastructure needed to capitalize on AI capabilities.	• **Guidelines & Regulation:** GDPR, Marketer Code of Conduct • **Transparency-Principle:** 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).	**Content creation with ChatGPT and Dall-E:** Generative AI tools are employed to create text and visual content for clients. AI is positioned as a competitive advantage for enhancing creativity and efficiency.	The organization focus on **punctuated equilibrium**, utilizing generative AI tools to optimize workflows and content as the current hype about ChatGPT is stressing the agencies daily work.	**#1 Standardization vs. personalized approaches:** The need to produce unique, high-value content to differentiate from what clients could 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?	• **Customer Experience**: Develop a strong customer-centric approach to create bespoke, high-quality experiences that differentiate the agency from generative AI-only 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	• **Data**: High availability of data, with ongoing progress toward	• **Regulation:** HR law, Car liability insurance	#1 **Damage assessment with AI**: Pilot project in the	The organization focuses on	#1 **Standardization vs. personalized approaches:**	• **Agile Culture:** Foster an agile workforce by implementing continuous

	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, though tempered by a cautious industrial culture.	establishing a centralized data management system. • **IT Resources**: Insufficient IT capacity to scale and integrate AI pilot projects into operational workflows effectively.	• **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:** Employees may feel devalued as their roles evolve due to AI integration (also customers could perceive the process as less empathetic).	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.	**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).	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). Reliance on external innovation brings challenges in maintaining transparency and fairness.	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.
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.	• **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.	**#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,	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	**#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	• **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.

	• **Culture**: Employees are accustomed to change management practices, fostering adaptability.		• **Non-maleficence-Principle:** Compliance with existing law must be ensured (e.g., European AI Act: labeling of contributions created by generative AI)	social media posts, and newsletters.	unsupervised) with generative AI tools (**exploration** strategies) when its about their marketing.	company's branding and emotional appeal. Over-reliance on AI might reduce creativity and critical thinking among employees.	• **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 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.	• **Data**: Machines are equipped with sensors for data tracking, but complexities and uncertainties in data ownership hinder further progress. • **IT Resources**: Limited IT infrastructure and skilled personnel, leading to reliance on external IT providers.	• **Regulation:** Struggle with different regulations as a multinational company (specifically Chinese vs. European law). • **Security-Principle:** Ethical concerns about data ownership and control. • **Non-maleficence-Principle:** Fear of replacement through disruptive AI players.	**AI detection in manufacturing:** Predicting machine wear and tear to prevent breakdowns and reduce production downtimes (predictive maintenance).	The organizations focus on **exploitation** as they fear to be replaced by disruptive players from other markets (especially China), as their regulation is different to the European ones, making it easier to stay competitive (**punctuated equilibrium**).	**Effective use of AI's capabilities vs. ethical violation**: Dependence on external AI platforms could expose the organization to risks such as data misuse or industry disruption by low-cost competitors.	• **Legal Requirements:** Advocate for state-level legal requirements to establish secure standards protecting against intellectual property theft and industry disruption. • **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	• **Data:** The organization resists tracking employee performance, focusing instead on open communication and acceptance, without monetary incentives**.**	• **Data security Principle:** In a hypothetical scenario, leadership should not have the right to access personal	**#1 Data analytics for performance evaluation**: The manager used data analytics to evaluate the performance and	The organization focuses on **exploration** strategies by designing innovative	#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.	• **Motivation to Learn**: Individuals should be motivated to understand and learn about algorithms to prevent overtrust in their outputs.

	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.	• **IT resources:** The organization's IT infrastructure is underdeveloped, largely due to management's reluctance toward AI adoption.	data for controlling employees, as this could violate privacy rights and undermine trust within the organization.	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.	technologies like the digital skiing teacher, aiming to experiment with new concepts and push the boundaries of what is possible in the field.	**#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.	• **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	• **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 prefeed 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.

	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.						• **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, Electronic Provider)	• **Strategic alignment**: The organization positions itself as a pioneer in AI adoption 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.	• **Data**: Machines are equipped with sensors for real-time data tracking; however, 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.	• **Sustainability-Principle**: Balancing technological advancements with sustainable resource utilization and environmental responsibility.	**Smart home:** Incorporating sensors into washing machines to determine the appropriate amount of water usage based on the weight of the clothes, enhancing efficiency and sustainability	The organization focuses on **exploration strategies**, using AI to develop new capabilities and enhance product innovation.	**Standardization vs. personalized approaches:** 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).	• **AI Training**: Promote transparency and user education about the benefits and limitations 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 streamline communication processes. • **Culture**: Embraces new technologies, works agilely, and stays consistently up-to-date with the latest advancements.	• **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 ease of use for all users, particularly those with limited technological experience.	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 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	• **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	• **Non-discrimination-Principle**: There is a critical issue that biases in underlying data models are reflected in	**Fraud detection in ticket sales**: The use case involves utilizing historical ticket sales data, including instances of fraudulent transactions, to train a	An organization that provides consulting services to other organizations on their technology products primarily	**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	• **AI Training:** Internal trainings on AI limits and the technologies ethical impacts to communicate it to customers • **AI Contracts**: Responsibilities must be contractually defined.

	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.	model storages for a great number of customers.	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.	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.	focuses on developing **exploitation** strategies to help their clients optimize processes and enhance productivity.	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.	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. • **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.	• **Data**: Relies on customers' data rather than owning proprietary datasets, focusing on data security and ethical usage. • **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.	• **Regulation:** Compliance with the European AI Act but facing gaps in industrial legal frameworks, particularly regarding accountability in 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	**AI in coding (GitHub Copilot) and application development (Open-AI):** Leveraging tools like GitHub Copilot and OpenAI Transformer Architecture to develop software applications. Focus on reducing manual coding effort, reusing existing code, and customizing applications for specific customer needs.	Focus on **exploration** strategies as a central component of the business model, emphasizing constant 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.	**#1 Standardization vs. personalized approaches:** Customers prioritize ease of use and intuitive AI models, whereas the organization must maintain critical knowledge of AI 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.	• **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 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

			complexities of AI mechanisms.				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, accountability, and a strong tolerance for failure. It embraces agile methodologies and encourages collaboration through cross-functional teams.	• **Data:** Own data management department for data analyses and forecast • **IT Resources:** Well-equipped organization (good data infrastructure and tools).	• **Sustainability-Principle:** Optim ize 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
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	• **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,	• **Data Security:** Some organizations are hesitant to adopt cloud services due to concerns about potential data loss.	**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,	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.	• **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

	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.	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.	• **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.	streamlining customer service operations.		**#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.	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 Manufacturer)	• **Strategic alignment**: The company prioritizes AI research and foster innovation. 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.	• **Data**: A dedicated data management department ensures high-quality data and 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.	• **Regulation:** Differences between Chinese law and European 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	**AI simulation and image generation**: Leveraging AI to create virtual rooms that facilitate interactive car design, offering a creative platform for innovation.	Focus on **exploration** as AI innovation is central to the 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,	**Effective use of AI's capabilities vs. ethical violation**: 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.	• **AI Guidelines:** Develop and implement a comprehensive internal ethics guide to outline 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

	• **Culture**: Organizational culture varies significantly; some departments demonstrate openness to innovation, while others maintain a more traditional, risk-averse stance.		property is critical, ensuring that sensitive information such as construction designs is not exposed through AI tools.		illustrating some levels of **exploitation**.		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 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.	• **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 not inadvertently harm organizational assets or escalate security vulnerabilities.	**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 constant evolution of the defensive	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 cybersecurity (**punctuated equilibrium**). The strategy is reactive, focused on using AI for immediate defense rather than proactively exploring AI's potential for broader applications.	**#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 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?	• **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.
E4 (Recruiter,	• **Strategic alignment**: The organization	• **Data**: Candidate data is managed systematically	• **Non-maleficence-**	**Candidate assessment with AI:**	The company employs AI	**#1 Effective use of AI's capabilities vs. ethical violation**:	• **Human-in-the-Loop:** For high-level positions (e.g.,

Headhunting Agency)	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)	but largely processed manually, limiting efficiency and scalability. • **IT Resources**: No dedicated IT infrastructure for AI adoption, which constrains the implementation of advanced tools.	**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.	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.	experimentally, focusing on identifying areas where it can enhance decision-making without compromising quality (rather **exploration**).	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.	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, Tax and Audit Consulting)	• **Strategic alignment**: The organization is in the exploratory phase, 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	• **Data**: Manages a diverse range of customer databases; 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.	• **Regulation:** GDPR compliance for 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.	**#1 Algorithmic platform for predictions:** 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	Focus on **exploration**, experimenting with the latest AI technologies to maintain market competitiveness and drive innovation.	**#1 Standardization vs. personalized approaches:** 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.	• **Ethical Training:** Educate employees and customers on the ethical implications of AI use, 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.

	adopting cutting-edge technology.			Microsoft Copilot for efficient coding, programming, and automation of repetitive tasks, enhancing developer productivity.			
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. • **Culture**: The company has a sedate, safety-oriented corporate culture, typical of the insurance industry (focus on risk mitigation).	• **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 tools that may erode creativity or personalization.	#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 creation with generative AI:** Leveraging ChatGPT and Dall-E for tasks like creating job offers and creating social media posts to enhance employer branding.	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 rather linked to **exploration** due to the medial hype surrounding AI (particularly ChatGPT; **punctuated equilibrium**).	**#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 ability to stand out, especially as competitors adopt similar tools.	• **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.
E7 (Recruiter, Technology Provider)	• **Strategic alignment**: AI is primarily used in recruitment processes and is mandatory for certain roles, focusing on optimizing hiring practices.	• **Data**: Data management practices require improvement. • **IT Resources**: Focuses on innovation for customer-facing products but lags in adopting modern	• **Fairness-Principle:** Aimed at reducing human bias but not completely bias-free**.** • **Transparency-Principle:** Candidates are	**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	The organization adopts a dual approach to AI implementation, experimenting with AI tools internally—such as in	**#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	• **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.

	• **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.	systems internally for processes like recruitment.	informed about AI use and must consent. • **Non-maleficence-Principle:** Concerns about the dehumanization of processes, especially for empathy-driven roles like porters.	personality tests and ranking candidates.	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**).	may differ significantly from in-person evaluations, leading to mixed perceptions of AI's reliability among both recruiters and candidates.	
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,	• **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	**#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	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

	fostering a culture of innovation and exploration.		nationalities when evaluating criminality, so users must critically reflect on potential biases and ensure fairness in the system.	collaborate in tasks such as assembly, welding, or material handling, improving productivity and safety.			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 knowledge of technology, presenting a challenge for integrating AI into their learning experience.	• **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 issues by enabling students to critically analyze and interpret AI decision-making processes.	**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 understanding of model transparency, and on how to develop models according to law.	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**: 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.	• **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. • **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	• **Data**: Does not own proprietary datasets but possesses strong expertise in managing	• **Regulation:** Lacks comprehensive legal frameworks to define	**Customer chatbot**: Pilot implementation of a chatbot for 1st-level customer support, handling	The organization focus on **exploration** strategies for	**Shared responsibility vs. clear allocation of individual responsibility:** The organization has deep knowledge of AI limitations (e.g.,	• **AI Training**: Provide training in prompt engineering and effective interaction with AI systems

	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.	and architecting data systems. • **IT Resources**: Combines in-house tools with open-source platforms to maximize flexibility and innovation.	accountability and liability in AI use. • **Accountability-Principle:** Responsibility for AI-related outcomes should primarily rest with those who develop and provide the tools.	tasks such as communication, newsletter dissemination, order confirmations, etc. (built on GPT-based architectures).	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.	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).	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 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.	• **Resources:** No details on IT resources at the department, however, the personnel had a shared understanding of AI and their respective role to incorporate their AI definition into the work with organizations. • **Data:** There are differing understandings between the organization and its providers regarding data quality and flow.	• **Transparency-Principle:** When AI functions as a „black box," it hinders the clear allocation of responsibility, making it difficult to trace decisions and hold individuals accountable.	**Anomaly Detection & Intelligent Manufacturing:** Use AI to monitor machines in real-time, detecting anomalies to prevent damage such as cracking, ensuring smooth operations and reducing downtime.	The organization focuses on **exploration** strategies by offering a new-to-market analytical tool for predictive maintenance, aiming to innovate and improve maintenance processes through advanced data-driven insights.	**#1 Trust vs. Overtrust:** Different understandings of AI lead to varied usage: some focus solely on its advantages or 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	• **Human Oversight:** Humans must remain in the loop to prevent incorrect AI performance and ensure accurate decision-making. • **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.

	• **Culture:** The organization is open to new technologies and actively explores research on enhancing digital competencies.					long-term reliance on external support.	• **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. However, the overall technological understanding among staff remains limited, with a preference for incremental innovation.	• **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 matching process and increases efficiency by automating the collection of customers' preferences and aligning them with suitable housing options.	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 content creation, emphasizing efficiency and incremental innovation over exploratory approaches.	**#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, organizations must reflect on whether the AI's actions and outcomes align with their moral values and ethical standards.	• **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. • **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 is responsible in case of an accident, but it should be ensured that humans can intervene if necessary to maintain control and accountability.	• **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.

Appendix II. Overview of All Works

Journal Articles

Published

Leffrang, D., Passlack, N., Müller, O., & Posegga, O. (2026). Beneficial Mistrust in Generative AI? The Role of AI Literacy in Programmers' Handling of Bad Coding Advice. *Behaviour and Information Technology*. https://doi.org/10.1007/s12599-026-01000-5.

Passlack, N., Hammerschmidt, T., & Posegga, O. (2025). With Great Power Comes Great Responsibility: What Shapes AI Literacy for Responsible Interactions of Knowledge Workers With AI?. *Information Systems Frontiers*. https://doi.org/10.1007/s10796–025–10648–5.

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*, 1–87. https://doi.org/10.1007/s10796–025–10596–0.

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.

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.

In progress

Passlack, N. (Working Paper, unpublished). From Teachers to Cheerleaders? Teachers' Role Responsibilities in the Age of Generative AI.

Hammerschmidt, T., Passlack, N., & Posegga, O. (Working Paper, unpublished). How Organizations Ethically Respond to Decision Dilemmas in AI Deployment.

Gerholz, K. & Passlack, N. (Working Paper, unpublished). Generative Artificial Intelligence (GAI) as Sparring Partner: Insights Into Teachers' Perceptions of Personal Learning Developments Through GAI Use.

Conference Proceedings

Published

Passlack, N., Hammerschmidt, T., & Posegga, O. (2025). How Leaders' Morality Influences Their Responsibility Perceptions When Relying on Algorithmic Advice. *AMCIS 2025 Proceedings*. 4. https://aisel.aisnet.org/amcis2025/is_leader/4/.

Passlack, N., Gerholz, K. H., & Schlottmann, P. (2025). Reskilling Me Softly: Perceived Changes in Students' Skilling When Using GenAI in Academic Research Projects. *Proceedings of the 58th Hawaii International Conference on System Sciences*. https://hdl.handle.net/10125/109440.

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*. https://hdl.handle.net/10125/106426.

Passlack, N. K., Heyder, T., Klemm, F., & Posegga, O. (2023). How Human–AI Collaboration Affects Attribution of Responsibility for Failure and Success. *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*. 15. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/15/.

Passlack, N. (2023). Alone together: Organizational measures to address pitfalls of virtual collaboration. *Proceedings of the 56th Hawaii International Conference on System Sciences*. https://aisel.aisnet.org/hicss-56/cl/distributed_collaboration/4.

Others

Published

Gerholz, K. & Passlack, N. (2025). Skillen statt Chillen!: Re–Skillingpotentiale von Lehrkräften durch die Nutzung von KI–Anwendungen. In *Berufs– und Wirtschaftspädagogik Online : bwp@, Hamburg: Institut für Berufs– und Wirtschaftspädagogik Universität Hamburg*, p. 1–17.

Gerholz, K. H., Leicht, S., Passlack, N., Schlottmann, P., & Wagner, A. (2025). Nutzen und Erleben der Berufseinstiegsbegleitung in Bayern (NEBerEbBay): Abschlussbericht 2024 In *Forschungsforum Berufs– und Wirtschaftspädagogik*.