Methods, Mechanisms, and Effects of Explainable AI-based Decision Support

Felix Haag

Bamberg, 2026

Diese Arbeit hat der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Erstgutachter: Prof. Dr. Thorsten Staake

Zweitgutachter: Prof. Dr. Patrick Zschech

Mitglied der Promotionskommission: Prof. Dr. Oliver Posegga

Tag der Disputation: 08.05.2026

URN: urn:nbn:de:bvb:473-irb-115736x

DOI: https://doi.org/10.20378/irb-115736

Acknowledgments

Without a doubt, these have been the most challenging, but also the most exciting years of my life. That I made it here is something I owe to the people who supported me along the way. Your support meant more to me than any words can fully express. But here, I will try:

First and foremost, I want to thank my supervisor, Prof. Dr. Thorsten Staake. There are many reasons I could thank him, above all for the constant feedback, guidance, and mentorship. But what I will remember most fondly are our research discussions: standing wide-eyed around the whiteboard, wholly consumed by whatever we had just thought of. I also owe a sincere thank-you to Prof. Dr. Patrick Zschech for his invaluable feedback, his shared enthusiasm for interpretable ML and XAI research, and his willingness to serve on my PhD committee. Furthermore, I would like to express my gratitude to Prof. Dr. Oliver Posegga for his thoughtful feedback on the theoretical embedding of my work and for chairing my PhD committee.

I also owe much to those whose guidance left a lasting mark on my work. Special thanks go to my co-supervisor, Prof. Dr. Konstantin Hopf, who introduced me to the world of IS research, always had an open ear whenever I needed advice, and – most importantly – encouraged me to pursue a PhD in the first place. Likewise, I want to thank Dr. Sebastian Günther, from whom I learned immensely and whose expertise in field studies, statistics, and writing left a lasting impression on me. Also, I would like to express my gratitude to Prof. Dr. Elgar Fleisch and Prof. Dr. Felix Wortmann for an exceptional stay at the University of St. Gallen.

Moreover, I am thankful to my colleagues who made the day-to-day work genuinely fun. First among them is my lab buddy, Dr. Carlo Stingl, with whom no day passed without a conversation, whether about research, life, or everything in between. Special thanks also go to Dr. Andreas Weigert, Christian Lorenz, Christian Weigl, David Schaurecker, Dr. Kevin O'Sullivan, Laura Schneider, Dr. Leonard Michels, Dr. Markus Kreft, Maximilian May, Dr. Prakhar Mehta, Samuel Schöb, Sarah Betz, Sophie Kuhlemann, and Dr. Tobias Brudermüller for the good times and lively exchanges. I would also like to thank the student research assistants Katrin Zerfass, Nasima Fakir, and Pedro Menelau Vasconcelos for their excellent research support. Looking back, I feel lucky to have met every single one of you.

Behind this dissertation are people without whom it would never have been written. To my friends, especially Carmen Herrmann, Jonas Troles, and Tobias Weiss, and the many others whose friendship I deeply cherish: thank you for always being there, even in moments of doubt. I am grateful to have you in my life. To my grandparents Else and Günther, my brother Max, and my parents Ute and Horst: thank you for your support and love throughout all these years.

Felix Haag
Bamberg, June 2026

Zusammenfassung (German summary)

Entwicklungen im Bereich der Künstlichen Intelligenz (KI) haben die Bearbeitung vieler Aufgaben in unserem Alltag und Berufsleben verändert. Empirische Studien zeigen, dass KI-Systeme in spezifischen Anwendungsfeldern Aufgaben automatisiert mit hoher Genauigkeit und teils überlegener Performanz im Vergleich zu vielen menschlichen Experten ausführen können (z. B. Esteva et al., 2017; Silver et al., 2016). Trotz beeindruckender Leistungen der KI verbleibt die Verantwortung für die Konsequenzen von KI-basierten Entscheidungen beim Menschen (Sturm et al., 2023). Damit verschiebt sich der Fokus für viele Aufgaben von der automatisierten Aufgabenausführung hin zu einer kollaborativen Interaktion zwischen Mensch und KI-basierter Entscheidungsunterstützung (Wilson und Daugherty, 2018).

Ein zentraler Erfolgsfaktor für diese Kollaboration scheint die Erklärbarkeit der zugrunde liegenden Modelle des maschinellen Lernens zu sein (vgl. Hemmer et al., 2022), die im Zentrum des Forschungsfelds der erklärbaren KI steht (vgl. Adadi und Berrada, 2018). Studien in der Literatur zu Informationssystemen (IS) verweisen auf Vorteile durch Erklärungen, wie beispielsweise ein höheres Vertrauen in KI (z. B. Meske und Bunde, 2022), eine höhere Akzeptanz (z. B. Gönül et al., 2006) sowie eine verbesserte menschliche Entscheidungsperformanz (z. B. Förster et al., 2025). Wobei sich in der Literatur ein relativ konsistentes Bild hinsichtlich des Einflusses von Erklärbarkeit auf Vertrauen zeigt (vgl. Atf und Lewis, 2025), ist das Bild zur Entscheidungsperformanz ambivalent: Während einige Studien positive Effekte durch erklärbare KI-basierte Entscheidungsunterstützung berichten (z. B. Lai et al., 2020), finden andere keine oder sogar negative Effekte (z. B. Bauer et al., 2023; Carton et al., 2020). Diese Heterogenität der Effekte legt nahe, dass die Effektivität von erklärbarer KI (d. h. die Wirkung von Erklärbarkeit als Merkmal von KI) für die Entscheidungsunterstützung von verschiedenen Einflussfaktoren abhängt – und weniger robust und konsistent ist, als häufig angenommen (vgl. Schemmer et al., 2022). Für die Diskrepanz zwischen der angenommenen und beobachteten Effektivität könnten mehrere Faktoren eine entscheidende Rolle spielen. Einerseits könnten Einschränkungen der derzeitigen Methoden, etwa in Bezug auf die Qualität der bereitgestellten Erklärungen, von Bedeutung sein (vgl. Rudin, 2019). Andererseits könnten menschliche Faktoren, wie Präferenzen oder kognitive Prozesse bei der Nutzung von Erklärungen, maßgeblich zur Heterogenität der Effekte beitragen (vgl. Bauer et al., 2021; Fürnkranz et al., 2020). Zudem wurden die Effekte von erklärbarer KI-basierter Entscheidungsunterstützung vor allem in kontrollierten Umgebungen untersucht (z. B. Förster et al., 2025), wohingegen Feldstudien, die reale Anwendungskontexte berücksichtigen, bislang selten sind (Brasse et al., 2023).

Vor diesem Hintergrund wird deutlich, dass Erklärbarkeit zwar häufig als Schlüsselfaktor für die Effektivität von KI-basierter Entscheidungsunterstützung angeführt wird (z. B. Hemmer et al., 2022), ihr tatsächlicher Beitrag zur Mensch-KI-Kollaboration in großen Teilen bislang jedoch unzureichend verstanden ist. Entsprechende Erkenntnisse könnten daher sowohl das

theoretische Fundament für die effektive Interaktion zwischen Mensch und erklärbarer KI erweitern als auch praktische Implikationen für die Gestaltung von erklärbaren KI-basierten Entscheidungsunterstützungssystemen aufzeigen.

Diese Dissertation adressiert diesen Forschungsbedarf, indem sie Methoden, Mechanismen und Effekte erklärbarer KI-basierter Entscheidungsunterstützung untersucht. Dazu (i) synthetisiert sie die bisherigen Befunde zum Effekt von erklärbarer KI-basierter Entscheidungsunterstützung auf die menschliche Entscheidungsperformanz, (ii) analysiert die zugrunde liegenden Mechanismen, mit welchen Erklärungen diese Performanz potenziell beeinflussen und (iii) ergänzt diese Arbeiten durch die Untersuchung eines erklärbare KI-basierten Entscheidungsunterstützungssystems im Feld. Damit leistet diese Dissertation einerseits theoretische Beiträge zur Interaktion zwischen Mensch und erklärbarer KI und zeigt andererseits praktische Implikationen, etwa zur Gestaltung von erklärbare KI-basierten Entscheidungsunterstützungssystemen, auf. Die Arbeiten stützen sich auf Anwendungen in unterschiedlichen Domänen, darunter die Energiebranche und digitale Hochschullehre. Insgesamt umfasst die Dissertation ein Rahmenpapier sowie neun wissenschaftliche Artikel, die in drei Kapiteln gebündelt sind.

Kapitel 1 schafft zunächst ein grundlegendes Verständnis für den Gesamteffekt von erklärbarer KI-basierter Entscheidungsunterstützung auf die menschliche Entscheidungsperformanz, indem es die Befunde der Literatur in einer Meta-Analyse synthetisiert. Die Ergebnisse bestätigen die zuvor postulierte Heterogenität: Die berichteten Effekte variieren deutlich zwischen den Studien. Zudem zeigt die Analyse, dass der über alle Studien hinweg zusammengefasste zusätzliche Effekt von Erklärungen im Vergleich zu reiner KI-Unterstützung zwar positiv, aber klein ist. Eine Moderationsanalyse zum Risiko des Bias in den berichteten Studienergebnissen zeigt darüber hinaus, dass ein höheres Bias-Risiko systematisch mit größeren berichteten Effekten assoziiert ist. Dies legt nahe, dass die tatsächliche Effektgröße noch kleiner ist, was die Effektivität erklärbarer KI-basierter Entscheidungsunterstützung zusätzlich infrage stellt. Darüber hinaus prüft das Kapitel den Erklärungstyp als moderierenden Faktor, der in der Literatur als wichtig angeführt wird (Herm, 2023) und potenziell Unterschiede im Effekt von erklärbare KI-basierter Entscheidungsunterstützung erklären könnte. Die Ergebnisse zeigen, dass der Erklärungstyp allein die Variabilität der Effekte zwischen den Studien nicht erklären kann. Damit konsolidiert das Kapitel die Ergebnisse der Literatur, bestätigt die Heterogenität der Befunde und motiviert die vertiefte Untersuchung von Mechanismen in domänenspezifischen Kontexten.

Kapitel 2 widmet sich detaillierten, domänenspezifischen Untersuchungen von potenziellen Mechanismen zur Erklärung von Unterschieden in der Effektivität erklärbarer KI-basierter Entscheidungsunterstützung. Das Kapitel betrachtet diese Mechanismen aus zwei Perspektiven: Einer methodenzentrierten und einer menschenzentrierten. Am Beispiel einer *Cross-Selling*-Vorhersage zu Produkten eines Energieversorgers zeigt das Kapitel, dass die Erklärungen einer erklärbaren KI-Methode für jenen Fall eine zeitliche Beständigkeit aufweisen. Darüber hinaus lassen die Ergebnisse erkennen, dass gängige post-hoc-Methoden in ihrer Erklärungsgüte zwar valide, jedoch begrenzt sind. Die Befunde der menschenzentrierten Perspektive legen nahe,

dass Unterschiede in der Verständlichkeit und Präferenz von Visualisierungen Mechanismen darstellen, die Unterschiede in der Entscheidungsperformanz erklären können. Zudem deuten die Ergebnisse darauf hin, dass die Effektivität von Erklärungen auch von ihrer Salienz abhängt, indem diese die Aufmerksamkeit der Nutzenden lenkt. Die Entscheidungsperformanz verbesserte sich durch eine Erklärung marginal statistisch signifikant nur dann, wenn kein starker externer Reiz (hier: numerischer Anker) dominierte. Darüber hinaus zeigen die Befunde, dass sowohl KI-gestützte als auch durch Erklärungen ergänzte Unterstützung den negativen Einfluss des Anker-Bias abschwächen kann. Damit liefert das Kapitel empirische Evidenz für eine Proposition, die auf dem *Selective Accessibility Model*, einem theoretischen Modell zur Erklärung des Anker-Bias-Phänomens, basiert (vgl. Mussweiler et al., 2000). Zusammengefasst identifiziert Kapitel 2 Mechanismen für die Effektivität erklärbarer KI-basierter Entscheidungsunterstützung und erweitert damit das theoretische Fundament für eine effektive Interaktion zwischen Mensch und erklärbarer KI.

Kapitel 3 erweitert die bislang seltene Evidenz aus Feldexperimenten (vgl. Brasse et al., 2023) und geht über bestehende Forschungsarbeiten hinaus, indem es ein neuartiges erklärbare KI-basiertes Entscheidungsunterstützungssystem instanziiert. Das System basiert auf Erkenntnissen der Theorie des selbstregulierten Lernens sowie auf kontrafaktischen Erklärungen aus der erklärbaren KI. Ziel der Instanziierung ist es, die Heterogenität der Lernenden bei der Feedbackbereitstellung zu berücksichtigen. Mithilfe auf historischen Daten trainierter Modelle generiert das System personalisiertes Feedback und schlägt Lernaktionen zur Verbesserung des prognostizierten Klausurergebnisses vor. Die Evaluation in einem Feldexperiment zeigt, dass das System die Lernergebnisse der Studierenden verbessert: Während der Interventionsphase verbrachten die Teilnehmenden der Treatmentgruppe pro Woche im Durchschnitt 80,7 % mehr Zeit in der digitalen Lernumgebung als die Kontrollgruppe. Die Treatmentgruppe erzielte zudem 13,2 Punkte (von 90 möglichen Punkten) mehr in der Klausur. Darüber hinaus profitieren Lernende – anders als in vielen anderen Studien (z. B. Leung et al., 2023; Mousavi et al., 2025) – unabhängig von Lernercharakteristika, die mit selbstreguliertem Lernen zusammenhängen (Metakognition, Prokrastination). Zusammenfassend zeigt Kapitel 3 einen Weg für den effektiven Einsatz erklärbarer KI-basierter Entscheidungsunterstützung in einem realen Anwendungskontext. Es verbindet dabei die in der Theorie des selbstregulierten Lernens beschriebene Heterogenität von Lernenden mit der Instanziierung adaptiver IS, die diese Unterschiede adressieren. Damit legt Kapitel 3 die Grundlage für eine neue Form von Entscheidungsunterstützungssystemen, die erklärbare KI einsetzt, um personalisiertes Feedback bereitzustellen.

Insgesamt leistet diese Dissertation einen Beitrag zum Verständnis der Methoden, Mechanismen und Effekte von erklärbarer KI-basierter Entscheidungsunterstützung. Durch die Identifikation von Mechanismen erweitert sie das theoretische Fundament für die Interaktion zwischen Mensch und erklärbarer KI. Zugleich bilden die Untersuchungen eine Grundlage für die effektive Gestaltung erklärbarer KI-basierter Entscheidungsunterstützungssysteme, die bei wichtigen Aufgaben im Alltag und im Berufsleben wirksam unterstützen können.

Literaturverzeichnis

Adadi, A. und Berrada, M. (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". *IEEE Access* 6, S. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

Atf, Z. und Lewis, P. R. (2025). "Is Trust Correlated With Explainability in AI? A Meta-Analysis". *IEEE Transactions on Technology and Society*, S. 1–8. DOI: 10.1109/TTS.2025.3558448.

Bauer, K., Hinz, O., van der Aalst, W. und Weinhardt, C. (2021). "Expl(AI)n It to Me – Explainable AI and Information Systems Research". *Business & Information Systems Engineering* 63 (2), S. 79–82. DOI: 10.1007/s12599-021-00683-2.

Bauer, K., Von Zahn, M. und Hinz, O. (2023). "Expl(AI)Ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing". *Information Systems Research* 34 (4), S. 1582–1602. DOI: 10.1287/isre.2023.1199.

Brasse, J., Broder, H. R., Förster, M., Klier, M. und Sigler, I. (2023). "Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions". *Electronic Markets* 33. DOI: 10.1007/s12525-023-00644-5.

Carton, S., Mei, Q. und Resnick, P. (2020). "Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity". *Proceedings of the 14th International AAAI Conference on Web and Social Media*. Atlanta, Georgia, USA (virtual). DOI: 10.1609/icwsm.v14i1.7282.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. und Thrun, S. (2017). "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks". *Nature* 542, S. 115–118. DOI: 10.1038/nature21056.

Förster, M., Broder, H. R., Fahr, M. C., Klier, M. und Fink, L. (2025). "Tell Me More, Tell Me More: The Impact of Explanations on Learning from Feedback Provided by Artificial Intelligence". *European Journal of Information Systems* 34 (2), S. 323–345. DOI: 10.1080/0960085X.2024.2404028.

Fürnkranz, J., Kliegr, T. und Paulheim, H. (2020). "On Cognitive Preferences and the Plausibility of Rule-Based Models". *Machine Learning* 109, S. 853–898. DOI: 10.1007/s10994-019-05856-5.

Gönül, M. S., Önkal, D. und Lawrence, M. (2006). "The Effects of Structural Characteristics of Explanations on Use of a DSS". *Decision Support Systems* 42 (3), S. 1481–1493. DOI: 10.1016/j.dss.2005.12.003.

Hemmer, P., Schemmer, M., Riefle, L., Rosellen, N., Vössing, M. und Kühl, N. (2022). "Factors That Influence the Adoption of Human-AI Collaboration in Clinical Decision-Making". *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Herm, L.-V. (2023). "Impact of Explainable AI on Cognitive Load: Insights from an Empirical Study". *Proceedings of the 31st European Conference on Information Systems*. Kristiansand, Norway.

Lai, V., Liu, H. und Tan, C. (2020). "'Why Is "Chicago" Deceptive?' Towards Building Model-Driven Tutorials for Humans". *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA. DOI: 10.1145/3313831.3376873.

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W. und Yue, W. T. (2023). "Could Gamification Designs Enhance Online Learning Through Personalization? Lessons from a Field Experiment". *Information Systems Research* 34 (1), S. 27–49. DOI: 10.1287/isre.2022.1123.

Meske, C. und Bunde, E. (2022). "Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection". *Information Systems Frontiers*. DOI: 10.1007/s10796-021-10234-5.

Mousavi, N., Golara, S. und Bockstedt, J. (2025). "The Impact of Situational Achievement Goals on Online Learning Behavior: Results from Field Experiments". *Information Systems Research* 36 (2), S. 983–1010. DOI: 10.1287/isre.2022.0353.

Mussweiler, T., Strack, F. und Pfeiffer, T. (2000). "Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility". en. *Personality and Social Psychology Bulletin* 26 (9), S. 1142–1150. DOI: 10.1177/01461672002611010.

Rudin, C. (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". *Nature Machine Intelligence* 1 (5), S. 206–215. DOI: 10.1038/s42256-019-0048-x.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N. und Vössing, M. (2022). "A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making". *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. Oxford, United Kingdom. DOI: 10.1145/3514094.3534128.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. und Hassabis, D. (2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature* 529, S. 484–489. DOI: 10.1038/nature16961.

Sturm, T., Pumplun, L., Gerlach, J. P., Kowalczyk, M. und Buxmann, P. (2023). "Machine Learning Advice in Managerial Decision-Making: The Overlooked Role of Decision Makers' Advice Utilization". *The Journal of Strategic Information Systems* 32 (4). DOI: 10.1016/j.jsis.2023.101790.

Wilson, H. J. und Daugherty, P. R. (2018). "Collaborative Intelligence: Humans and AI Are Joining Forces". *Harvard Business Review* 96, S. 114–123.

Table of contents

Introductory paper

Table of contents

1 Introduction

Recent developments in artificial intelligence (AI) have led to notable cases where it matches – and in some instances surpasses – the performance of human experts (see, e.g., Esteva et al., 2017; Silver et al., 2016). Consequently, AI has begun to transform many aspects of daily life, for example, by automating decision-making processes in labor-related routines and tasks (Agrawal et al., 2023). Such takeovers can fuel fears about diminishing control and technological singularity (Bostrom, 2014) but also open up new opportunities for productivity gains (Agrawal et al., 2018). Yet in many situations, humans are reluctant to delegate decisions entirely to algorithms (Burton et al., 2020) as it is the human actors who ultimately remain accountable (Sturm et al., 2023). This tension has shifted the focus for numerous tasks from replacement to collaboration by combining human judgment and oversight with the strengths of AI. In these collaborative settings, AI acts as a decision aid, supporting rather than replacing human decision-making (Wilson and Daugherty, 2018).

A characteristic that appears critical for the adoption and effectiveness of decision support is the explainability of system outputs (Hemmer et al., 2022; Yeomans et al., 2019). Indeed, research in information systems (IS) and decision support systems (DSS) has demonstrated that such explanations can build trust (see, e.g., Wang and Benbasat, 2007) and, in some cases, enhance decision performance (see, e.g., Gregor and Benbasat, 1999). However, in AI-based DSS, the underlying support often uses machine learning (ML), the core technology of AI, which frequently involves models that are not inherently interpretable to humans (Coussement et al., 2024). This shortcoming can hinder the widespread adoption of AI and lead to distrust in the system (Shin, 2021; Wanner et al., 2020). The research field addressing this challenge is known as explainable artificial intelligence (XAI) and focuses on developing methods that explain ML outputs in human-understandable terms (Adadi and Berrada, 2018; Meske et al., 2020). The application of XAI has gained traction not only as a means to foster adoption but also as an additional source of information to improve human decision-making (see, e.g., Buçinca et al., 2020; van der Waa et al., 2021). In addition, IS and related research fields have explored the role of explanations in AI-based decision support (referred to as XAI-based decision support) and their effect on decision-making from a sociotechnical perspective (see, e.g., Herm, 2023b; Schemmer, Kühl, et al., 2022; Wang and Yin, 2021).

Employing explanations in IS is often portrayed as beneficial for human decision-making (Gregor and Benbasat, 1999; Wang and Benbasat, 2007), and some studies have indeed shown that XAI-based decision support can improve human task performance in certain contexts (see, e.g., Bansal et al., 2021; Lai et al., 2020). However, a closer examination of the broader empirical landscape of XAI studies that involve users uncovers a less promising picture: recent meta-analyses reveal that the overall effect of explanations in XAI-based decision support on task performance is negligibly small (Schemmer, Hemmer, et al., 2022). This suggests that

the presumed benefits of XAI for decision-making are far less robust and consistent than one might assume. In other words, while explanations in XAI-based decision support may appear promising, their overall effectiveness in reliably enhancing human task performance remains questionable.

To better understand the discrepancy between claimed and actual benefits, it is important to consider the potential mechanisms behind the inconsistent findings. On the one hand, these inconsistencies may stem from limitations in current XAI methods themselves, such as the overall quality of the explanations provided (Rudin, 2019). On the other hand, factors relevant to human-XAI interaction may play a substantial role, including user understanding and preferences, and more specifically, those related to human decision-making, such as cognitive processes and biases (Bauer et al., 2021; Fürnkranz et al., 2020; Schemmer, Kühl, et al., 2022). Thus, there is a need to gain a clearer understanding of these mechanisms and how both method- and human-related factors contribute to the inconsistencies in findings. In addition, this motivates the instantiation and evaluation of an XAI-based DSS that can provide adaptive, context-sensitive decision support tailored to individual users.

In light of this shortcoming, this dissertation aims to deepen understanding of the methods, mechanisms, and effects of XAI-based decision support from a sociotechnical perspective – that is, a perspective that considers both the technical aspects and the human factors associated with the use of IS (Sarker et al., 2019). Specifically, this dissertation validates the explanation quality of XAI methods (i.e., method-centric mechanisms), evaluates the overall effects of XAI-based decision support on decision-making and performance, and examines how human-related factors (i.e., human-centric mechanisms) shape the effectiveness of such support.

This overarching research endeavor is explored in greater detail and addressed through three research objectives. The first research objective focuses on establishing a basic understanding of the overall effect of XAI-based decision support on human decision-making performance. Specifically, it examines the existing inconsistency in reported results in the literature: while some studies report improvements in task performance when users rely on XAI-based decision support (see, e.g., Bansal et al., 2021; Lai et al., 2020), others find little to no or even negative effects when compared to purely AI-based decision support (see, e.g., Bauer et al., 2023; Carton et al., 2020). Although a previous meta-analysis based on nine articles identified an overall small and statistically insignificant effect of explanations in XAI-based decision support (Schemmer, Hemmer, et al., 2022), it is prudent to re-examine this finding using a broader and more up-to-date body of literature. In sum, this research objective addresses the overarching research need concerning the synthetization of the overall effect of XAI-based decision support on human task performance and thus lays the foundation for the subsequent research objectives. The first research objective is as follows:

> **Research Objective 1**: *To synthesize the overall effect of XAI-based decision support on human task performance across empirical studies in the literature.*

The second research objective focuses on assessing explanations in XAI-based decision support by examining mechanisms that may account for variation in its effectiveness. These mechanisms can be evaluated at two complementary levels: the method-centric perspective, which includes validation of XAI explanations, and the human-centric perspective, which concerns user-related factors such as comprehensibility and preferences in XAI visualizations, as well as cognitive biases, thereby reflecting a broader distinction in the literature between a technical focus and the study of human-XAI interaction (Schauer, 2024). The method-centric perspective in this dissertation specifically focuses on explanation quality and serves as a foundation for effective interaction with XAI-based decision support. In addition, the emphasis on human evaluation of explanations in such support addresses a critical shortcoming in the literature: only a small fraction of XAI research has empirically tested explanations with humans (Nauta et al., 2023; Suh et al., 2025). By combining both method-centric validation and human-centric evaluation, these efforts empirically contribute to the assessment of explanations in XAI-based decision support. Accordingly, the second research objective is as follows:

> **Research Objective 2**: *To validate and evaluate explanations in XAI-based decision support.*

The third research objective of this dissertation focuses on the instantiation of an XAI-based DSS and the evaluation of its effects in the field. In contrast to the previous research objective, which focused on explanations in XAI-based decision support within controlled environments, this objective aims to instantiate and test an XAI-based DSS with real users. In doing so, it complements the dissertation's research agenda by seeking to meaningfully integrate XAI as a system component within a DSS artifact to address a domain-specific challenge: providing personalized feedback in digital learning environments by accounting for differences among learners (Maier and Klotz, 2022). While the technical aspects of XAI have received increasing attention, its effects on humans in realistic scenarios – such as actual university courses embedded within digital learning environments – have largely been overlooked in the literature (Brasse et al., 2023). The third research objective is therefore the following:

> **Research Objective 3**: *To instantiate an XAI-based DSS that provides personalized feedback in digital learning environments and to evaluate its effects in the field.*

Taken together, this dissertation consists of the present introductory paper and nine accompanying research papers, which collectively address the three research objectives outlined above. In addition to providing the motivation and justification for the research objectives, the introductory paper covers the conceptual background and complementary theoretical foundations, the methodological approach, the main results, the contributions to the literature, and the implications for practice.

The papers are organized into three chapters, aligned with the three research objectives of this dissertation:

- **Chapter 1** synthesizes the overall effect (i.e., pooled effect) of XAI-based decision support on human task performance through a meta-analysis that combines effect sizes from multiple empirical studies (**Papers I**–**II**).
- **Chapter 2** validates and evaluates explanations in XAI-based decision support. Specifically, this chapter includes the method-centric validation of explanation quality and the human-centric evaluation of explanations with respect to user comprehensibility and preferences, and the potential mitigation of anchoring bias through XAI (**Papers III**–**VI**).
- **Chapter 3** first presents the instantiation of an XAI-based DSS that provides personalized feedback to learners. The subsequent evaluation of its effects takes place within an actual university course in a digital learning environment and is implemented as a field experiment (**Papers VII**–**IX**).

Figure 1 provides an overview of the structure of this dissertation by presenting the corresponding chapters, the included papers, and their specific focus.

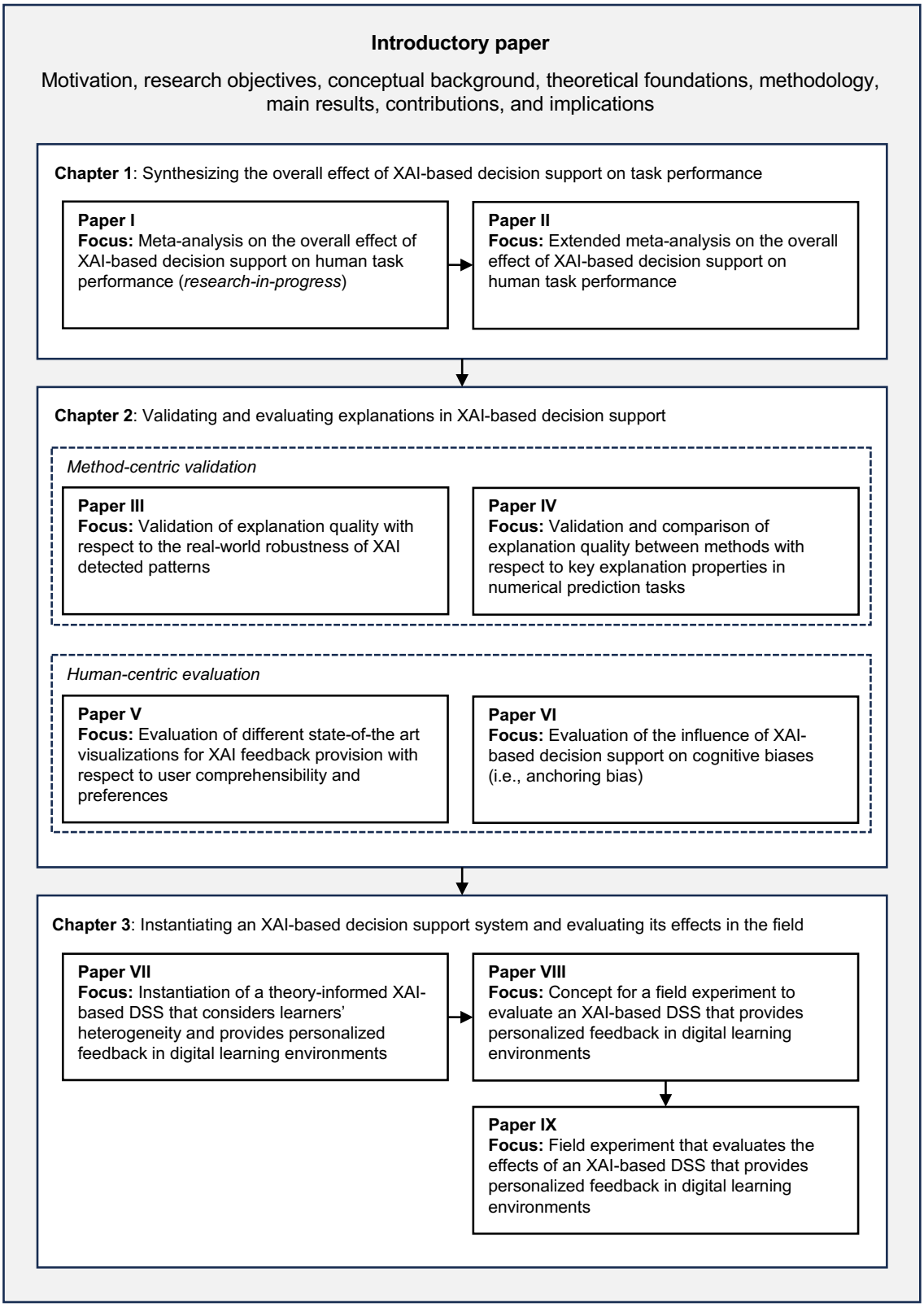


Figure 1: Dissertation structure

2 Conceptual background and theoretical foundations

This section introduces the conceptual background and domain-specific theoretical foundations that are relevant to the papers in this dissertation. It begins with an overview of the definitions and objectives of XAI. Next, it introduces two approaches relevant to the assessment of explanations, namely method-centric validation and human-centric evaluation, and outlines the key concepts associated with each. The section then presents the complementary theoretical foundations that underpin the experimental research component of this dissertation.

2.1 Definitions and objectives of explainable artificial intelligence (XAI)

The need to integrate explanations for the predictions of AI-based systems has been a long-term concern in AI and IS research. An early example in AI research is the expert system "MYCIN", developed in the 1970s to diagnose infectious diseases and that already included an explanation for its rule-based diagnoses (Buchanan and Shortliffe, 1984). In the 1980s, further research on explainability in expert systems appeared (see, e.g., Swartout, 1983), while in the 1990s seminal contributions proposed methods for the representation of neural networks (see, e.g., Craven and Shavlik, 1995). Shortly after, IS research explored system explainability by adopting a human-centric perspective and highlighting the positive effects of well-designed explanations in knowledge-based systems (Gregor and Benbasat, 1999). Consequently, embedding explainability into AI-based systems is not a new system requirement but a long-standing concern in both IS and AI research (Martens et al., 2025).

According to Adadi and Berrada (2018), the term XAI was first coined by Van Lent et al. (2004), who referred to XAI in the context of explaining AI agent behavior. Although research interest with respect to the term XAI has spiked in the last decade (Adadi and Berrada, 2018) – which some credit to the use of the term in a Defense Advanced Research Projects Agency (DARPA) program* (see, e.g., Martens et al., 2025) – a generally accepted definition of XAI and related terms has not yet been established (Bauer et al., 2023; Weber et al., 2024). Bauer et al. (2023) conceptualized XAI "as methods that possess the ability to present in understandable terms to a human why an AI system makes certain predictions" (p. 1584). Similarly, Adadi and Berrada (2018) defined XAI as "a research field that aims to make AI systems results more understandable to humans" (p. 52139). Taking inspiration from these authors, this dissertation defines XAI as a research field that promotes methods aimed at explaining the predictions of AI-based systems in human-understandable terms.

∗ See Gunning et al. (2021) for more details on the DARPA program.

An important distinction in this context regards interpretability versus explainability. Following Meske et al. (2020), this dissertation differentiates between these terms as follows: If a human can directly understand the predictions of a model without needing further explanation, this refers to interpretability and interpretable ML (Guidotti et al., 2019); interpretability can therefore be understood as a model characteristic (Rudin, 2019). In contrast, if understanding requires a proxy to interpret the potentially complex reasoning of a ML model, this refers to explainability and XAI (Adadi and Berrada, 2018). Given that this dissertation centers around methods that aim to explain the outputs of non-inherently interpretable ML models, it focuses on XAI and explainability. XAI-based explanations therefore refer to XAI method outputs and their efforts to make ML-based predictions understandable. Furthermore, in the context of this dissertation, the term "AI" refers to ML and ML-based predictions.

As reflected in the definition of XAI, the field of XAI research is inherently interdisciplinary. More specifically, Miller (2019) argued that XAI constitutes a human-agent interaction problem situated at the intersection of AI, social science, and human-computer interaction research. Furthermore, he highlighted that XAI research in the social sciences is closely tied to system design and user interaction.

The objectives one pursues with XAI seem to depend predominantly on the target audience (Meske et al., 2020). The objectives described by Meske et al. (2020) – some of which conceptually align with those of Adadi and Berrada (2018) – can be summarized as follows:

→ **Explainability to evaluate AI***: XAI may allow a deeper understanding of AI-based system behavior and the identification of potential weaknesses (Adadi and Berrada, 2018). This might benefit AI developers as it could help mitigate erroneous results, such as those caused by spurious correlations in ML models (Meske et al., 2020).

→ **Explainability to improve AI**: XAI can enhance and refine AI-based systems by revealing insights into their internal workings and therefore might support developers in improving accuracy (Meske et al., 2020). Furthermore, it could serve as a base for iterative improvements between humans and machines (Adadi and Berrada, 2018).

→ **Explainability to justify AI**: The use of AI-based systems may require justifications, especially for critical decisions with high stakes (Meske et al., 2020). XAI could help ensure that such decisions are not made erroneously and could provide reasoning for specific outcomes (Adadi and Berrada, 2018), thus assisting AI regulators (Meske et al., 2020). It may also help justify unexpected results, thus fostering trust in the system (Adadi and Berrada, 2018).

→ **Explainability to learn from AI***: XAI could enable users to derive new insights related to the problem or task at hand (Adadi and Berrada, 2018). More specifically, as

* The objectives *explainability to evaluate* and *explainability to learn* are described similarly by Adadi and Berrada (2018) as *explain to control* and *explain to discover*, respectively.

ML models can automatically identify patterns in data, XAI could also have the potential to reveal previously unknown insights (Meske et al., 2020). These insights can support human decision-making (see, e.g., Lai et al., 2020).

→ **Explainability to manage AI**: The adoption of AI can introduce new complexities to organizations by altering processes and routines (Meske et al., 2020). To address these challenges, Meske et al. (2020) argue that explainability is necessary for evaluating, improving, and justifying AI, as well as to learn from it (that is, achieving all the objectives above). Hence, XAI may be necessary to manage AI effectively.

Building on the concepts introduced above, this dissertation mainly focuses on explainability to learn from AI in the context of XAI-based decision support. Specifically, XAI-based decision support refers to AI-based decision support that is extended by explanations provided by XAI methods and aimed at enhancing human decision-making (i.e., *learn* in this context) (see, e.g., Buçinca et al., 2020; Herm, 2023b; Lai et al., 2020; Senoner et al., 2022; van der Waa et al., 2021). When this support is integrated into an IS artifact that provides recommendations and additional information to aid the decision-making process, it can be referred to as an XAI-based DSS (based on Herm, 2023a, referring to Arnott and Pervan, 2008).

2.2 Validation and evaluation approaches

In the context of IS research, validation can be defined as "checking the appropriateness of the system for the purpose for which it is being used" (Finlay, 1994, p. 209). Consequently, assessing the appropriateness of XAI methods for a subsequent integration in DSS refers in this dissertation to validation and, more specifically, method-centric validation. Evaluation, in contrast, can be described as the process of measuring "the benefit of the system to the users, project sponsors and ultimately the organization" (O'Keefe and O'Leary, 1993, p. 5). In this dissertation, evaluation therefore denotes assessing (i) the benefits of explanations in XAI-based decision support and (ii) XAI-based DSS, both through experiments involving humans.

Another categorization that has been widely adopted in XAI-related research (see, e.g., Brasse et al., 2023; Madsen et al., 2023) is the taxonomy provided by Doshi-Velez and Kim (2017) for assessing interpretability. This taxonomy divides approaches for assessment into three categories: functionally, human-, and application-grounded. Figure 2 illustrates how these categories relate to the validation and evaluation approaches adopted in this dissertation.

Functionally grounded is the least costly and most specific category, as it assesses explanation quality without involving humans (Doshi-Velez and Kim, 2017). Such assessments can therefore be less context-dependent (i.e., the user, task, and usage of explanations can remain undefined). Specifically, this category includes, for example, analyses that evaluate methods in terms of predictive performance while maintaining interpretability (Doshi-Velez and Kim, 2017). It also encompasses assessments related to method properties such as *faithfulness* (see, e.g.,

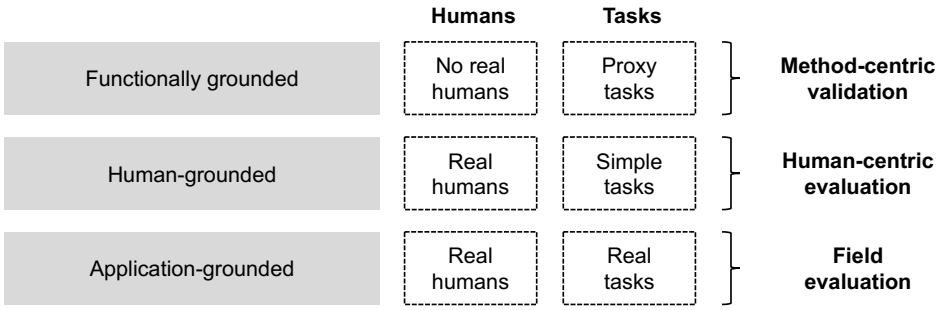


Figure 2: Mapping between the taxonomy of Doshi-Velez and Kim (2017) and the corresponding validation and evaluation approaches applied in this dissertation (illustration of the taxonomy adapted from Doshi-Velez and Kim, 2017)

Canha et al., 2025). In the context of this dissertation, this category refers to method-centric validation as it serves to assess the suitability of XAI methods in relation to the purpose for which they are used (i.e., explaining the outputs of ML models).

Human-grounded evaluations are characterized by Doshi-Velez and Kim (2017) as involving human experiments that comprise simple tasks (e.g., users choose preferred XAI visualizations). The rationale is that such experiments aim to evaluate explanations in relation to their general notions, for instance, within abstract tasks where many confounders (e.g., the task environment) can be controlled. The advantage of this type of evaluation is that scholars can rely on laypeople (e.g., through online labor platforms), enabling less costly evaluations and a larger pool of study participants. This is particularly useful when access to the target community is limited (e.g., in studies involving students constrained by semester schedules). In the context of this dissertation, human-grounded evaluations refer to human-centric evaluations within the scope of assessing explanations in XAI-based decision support.

Application-grounded evaluations involve human experiments in real application domains (Doshi-Velez and Kim, 2017). Subjects are thereby drawn from a domain-specific group relevant to the investigation (e.g., physicians diagnosing patients). Due to the complexity of such settings, these experiments are typically the most expensive and the most specific as they target a particular task within a given domain rather than the general notion of explanations (Doshi-Velez and Kim, 2017). In other words, such evaluations could be understood as placing more emphasis on the effect of the explanation or system on real-world outcomes than on the explanation itself. Hence, in the context of this dissertation, this denotes the evaluation of an XAI-based DSS in the field.

2.3 Method-centric validation

Building on the mapping above (see Figure 2), this subsection outlines concepts that are particularly relevant to the method-centric validation of explanations in XAI-based decision support (Paper III and Paper IV).

2.3.1 Method characteristics and explanation types

XAI research has produced a wide range of methods (see, e.g., Lundberg et al., 2020; Mothilal et al., 2020). Alongside this development, literature has outlined various characteristics to describe these methods and their resulting explanations, as well as to distinguish between them (see, e.g., Adadi and Berrada, 2018; Arrieta et al., 2020; Schwalbe and Finzel, 2023).

Three method characteristics are particularly relevant to this dissertation. First, *local and global interpretability* (Mohseni et al., 2021). Methods that allow the explanation of a prediction at the level of single observations offer local interpretability (Adadi and Berrada, 2018). When a method focuses on the model as a whole and can explain its overall logic and behavior that leads to all outcomes, it offers global interpretability (Adadi and Berrada, 2018; Meske et al., 2020). Second, *intrinsically interpretable models vs. post-hoc methods* (Adadi and Berrada, 2018). Intrinsically interpretable models are interpretable by design and do not require further explanation procedures. Post-hoc methods (also described by Mohseni et al., 2021 as "ad hoc explainers") are separate ex-post applied techniques that explain a model that is not inherently interpretable, such as a model considered a black-box (Adadi and Berrada, 2018). Third, *model-specific vs. model-agnostic methods* (Meske et al., 2020). Model-specific methods are limited to particular classes of models (e.g., kernel- or tree-based algorithms); intrinsically interpretable models are therefore model-specific by design (Adadi and Berrada, 2018). In contrast, model-agnostic approaches can be applied to any model as they are not tied to any specific class of models (Adadi and Berrada, 2018; Meske et al., 2020).

Another way to distinguish between methods is the *explanation type* that describes the type of explanatory information an explainer method provides (Herm, 2023b; Mohseni et al., 2021). Mohseni et al. (2021) outlined six explanation types relevant to the development of XAI-based systems (Table 1).

The papers in this dissertation that focus on method-centric validation address the explanation type *Why?* through feature attribution explanations, which describe how individual features contribute to a particular model prediction. Specifically, Paper III focuses on the XAI method Shapley additive explanations (SHAP), introduced by Lundberg and Lee (2017), due to its desirable properties for the paper's context, namely local interpretability and model-agnosticism. Paper IV considers multiple methods ranging from linear regression (i.e., intrinsically interpretable and model-specific) to model-agnostic post-hoc methods such as SHAP and local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016).

Table 1: XAI explanation types (based on Mohseni et al., 2021, and the descriptions of Herm, 2023b)

Explanation type	Description
Why?	Describe the specific input features or elements of the model's internal logic that led to the given prediction.
Why not?	Present contrastive explanations by describing why a model did not generate a specific prediction based on the input (i.e., help to understand the discrepancy between an actual prediction and a user's expected prediction).
What-else?	Provide input instances that yield identical or similar model predictions to those of the original input.
What-if?	Demonstrate how changes to the model (e.g., modifying model parameters) or the data (e.g., manipulating input features) affect the model's predictions.
How?	Provide an overall presentation of the model's functioning (i.e., its decision logic) to explain how the model operates.
How-to?	Describe how to shift the model's prediction toward a desired output by altering the input features or the model itself (i.e., a counterfactual explanation).

Beyond the method-centric validation component of this dissertation, explanation types are an important design-influencing factor in XAI-based systems (Mohseni et al., 2021) potentially affecting users' behavioral outcomes and task performance (Herm, 2023b). The impact of explanation types is therefore further analyzed in the meta-analysis of this dissertation (Paper I and Paper II).

2.3.2 Properties for explanation quality

The method-centric validation of XAI methods involves determining explanation quality based on proxy tasks (see Figure 2). From a practical perspective, such validations are particularly appealing because they are relatively inexpensive compared to evaluations with humans (Doshi-Velez and Kim, 2017). Moreover, they can rely on non-critical test data, which might be important when selecting an XAI method in a critical decision environment.

The literature defines a variety of properties and corresponding metrics to assess the quality of explanations (see, e.g., Schwalbe and Finzel, 2023). Among the various metrics that have been proposed, three frequently mentioned metrics are *consistency*, *stability*, and *fidelity* (see, e.g., Robnik-Šikonja and Bohanec, 2018; Schwalbe and Finzel, 2023).

Consistency describes the degree to which different models produce similar explanations when trained on the same task (i.e., predicting the same target variable) (Robnik-Šikonja and Bohanec, 2018). Although a set of different models may yield similar predictive performance, their internal structures, and consequently the explanations, can differ – a phenomenon known as the *Rashomon effect* in ML research (Breiman, 2001; Molnar, 2025; Rosenberger et al., 2025). Regarding *consistency*, Molnar (2025) pointed out that due to the Rashomon effect, it can be challenging to measure *consistency* as different models may use different features to make predictions, leading to inherently different explanations.

Stability is defined as the degree to which the explanations are similar for similar observations (Robnik-Šikonja and Bohanec, 2018). Unlike *consistency*, which considers variation in explanations across different models, *stability* focuses on variation within a single model when applied to similar inputs (Molnar, 2025). Here, explanations may also vary due to the use of different methods as they can rely on different approaches to derive explanations (Robnik-Šikonja and Bohanec, 2018). Among many approaches, stability can be measured through variations in feature values and resulting changes in the explanation (Molnar, 2025) or through the coherence of explanations for similar observations (see, e.g., Velmurugan et al., 2021).

Fidelity describes how well an explanation captures the behavior of a model (Robnik-Šikonja and Bohanec, 2018). Local *fidelity* refers to the extent to which an explanation approximates the model's behavior for a single prediction or a small subset of predictions (Molnar, 2025). However, local *fidelity* does not necessarily imply global *fidelity* as features may be important in a local context but not across the model's global behavior (Robnik-Šikonja and Bohanec, 2018). The terms *fidelity* and *faithfulness* are often used interchangeably (Schwalbe and Finzel, 2023); Paper IV and Paper V use the term "faithfulness". One common approach to measuring *faithfulness* involves perturbing important input features and evaluating the impact on the model's prediction (see, e.g., Alvarez Melis and Jaakkola, 2018; Du et al., 2019).

The concept of properties for explanation quality is particularly relevant for Paper III and Paper IV. Specifically, Paper III performs a robustness analysis that is similar to the definition of the metric *stability*, whereas Paper IV implements and measures the metrics *consistency*, *stability*, and *faithfulness*, with a focus on numerical prediction tasks.

2.4 Human-centric evaluation

Based on Figure 2, this section introduces key concepts that are particularly relevant for the human-centric evaluation of explanations in XAI-based decision support (Paper V and Paper VI).

2.4.1 Behavioral outcomes and task performance

XAI-based decision support can have multiple effects on users. In this dissertation, the effects investigated with respect to this support are (a) behavioral outcomes and (b) task performance. This subsection defines both outcomes, distinguishes between them, and explains their relationship within the context of this dissertation.

Behavioral outcomes refer to observable effects in human behavior that result from interactions with XAI-based decision support. Within the scope of this dissertation, these effects encompass, for example, altered price estimation behavior through explanations in AI-based decision support in the presence of an anchor (see Paper VI). In addition, outcomes extend to real-world behavioral changes, such as shifts in time spent on a digital learning platform or

changes in exam participation in higher education settings (see Paper IX). In general, behavioral outcomes refer, in the context of this dissertation, to users' behavioral responses to the exposure of XAI-based decision support.

Task performance, in turn, refers to an outcome that describes the success of the completion of a specific task when assisted through XAI-based decision support. Task performance can be measured in various ways, including the required time to complete a task (see, e.g., Lim et al., 2009) or decision accuracy (see, e.g., van der Waa et al., 2021). In the context of this dissertation, task performance is operationalized using three distinct metrics: accuracy in classification tasks (see Papers I, II, and V); the error in numerical estimation tasks (see Paper VI); and through students' exam performance in the context of higher education learning (see Paper IX). Accuracy in classification tasks and the error in numerical estimation tasks can be evaluated in a manner analogous to the predictive performance of ML models (see, e.g., Hastie et al., 2009).

Behavioral outcomes and task performance relate to each other in that behavioral outcomes can influence how successfully tasks are performed. More specifically, XAI-based decision support can shape behaviors that, in turn, have an impact on task performance. For example, XAI-based DSS in a digital learning platform may lead students to increase online learning time, which might contribute to an increase in task performance (e.g., improved exam performance).

2.4.2 User comprehensibility and preferences

User comprehensibility measures how well humans understand explanations (Molnar, 2025; Robnik-Šikonja and Bohanec, 2018). More specifically, comprehensibility can be defined as the degree to which users understand a model well enough to apply the acquired knowledge to new data and tasks (Fürnkranz et al., 2020). As users and tasks in the context of XAI research vary, there is some consensus in the literature that the comprehensibility of the explanation depends largely on the target audience and their objectives (see, e.g., Arrieta et al., 2020; Molnar, 2025). AI users, for example, could be interested in comparing the reasoning of a model with their own, while AI developers might be more concerned with optimizing models (Meske et al., 2020). In addition, the level of experience with AI (i.e., AI novices vs. AI experts) might also play a role in the human understanding of an explanation (Arrieta et al., 2020). In summary, the degree of comprehensibility seems to depend on the user's AI capabilities and objectives (see Section 2.1 for more details on the objectives of XAI).

Comprehensibility is influenced by various factors and can be difficult to measure (Molnar, 2025). Human-centric approaches to measurement include evaluating how well individuals can predict ML model behavior based on the explanation (Molnar, 2025). Another approach involves asking user-scenario-related questions based on the provided results; for example, Herm et al. (2021) outlined such an evaluation in the context of three domains ranging from low- to high-stakes decisions. Furthermore, there are approaches to measuring comprehensibility based

on the properties of an explanation (e.g., the depth and number of decision rules; see Verbeke et al., 2011), though these are beyond the scope of this dissertation.

Closely related to comprehensibility – but more concerned with the acceptance of explanations – are the cognitive preferences of users (Fürnkranz et al., 2020)*. User preferences are based on the subjective perception of individuals and refer to the perceived value of a particular explanation. Thus, an explanation can be highly comprehensible yet still be less preferred by users (Fürnkranz et al., 2020). Notably, Mohseni et al. (2021) listed the failure to consider user preferences when designing XAI-based systems as a pitfall. Consequently, user preferences might play a crucial role in the effectiveness of an XAI-based DSS with respect to desirable behavioral outcomes and task performance.

User comprehensibility and preferences are relevant to Paper V, which investigates both concepts with respect to different XAI visualizations.

2.5 Complementary theoretical foundations

After the more general introduction to key concepts, this subsection outlines the complementary theoretical foundations that inform and contextualize the dissertation's research. These theories help position the work within two specific domains: (a) XAI-based decision support in the presence of cognitive biases and (b) decision support for learning in higher education. To address anchoring bias in XAI-based decision support, the selective accessibility model provides a basis for strategies that counteract such biases (Paper VI). Additionally, self-regulated learning theory informs the instantiation and evaluation of an XAI-based DSS (Papers VII–IX).

2.5.1 Anchoring bias and the selective accessibility model

When employing XAI-based decision support, applications range from decisions that are situated in quick, intuitive judgments (e.g., quick food choices; see Buçinca et al., 2020), to those that require more deliberate, analytical reasoning (e.g., complex sales forecasts in an analytics dashboard; see Wang and Ding, 2024).

Building on the general distinction between decision characteristics, behavioral theorists have categorized decisions into two process modes (Evans, 2008): In system 1, humans make decisions in unconscious processes that involve automatic and implicit operations (i.e., fast decision-making including habitual behavior such as showering and washing hands; see, e.g., Stingl et al., 2022; Tiefenbeck et al., 2018). In system 2, decisions are driven by more controlled processes involving deliberate and serial operations (i.e., slow, analytical decision-making) (Kahneman, 2011). While system 1 helps us to act efficiently, it also often leads to taking shortcuts, ignoring valuable information, and falling prey to cognitive biases (Tversky and Kahneman, 1974). Cognitive biases can be defined as "Systematic error in judgment and decision-making common

* Fürnkranz et al. (2020) subsumed user acceptance and user preferences under the term *plausibility*.

to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments" (Wilke and Mata, 2012, p. 531).

Among the many cognitive biases, one of the most robust is anchoring bias (Furnham and Boo, 2011), which refers to the influence of an initial piece of information (the anchor) on subsequent decision-making (Tversky and Kahneman, 1974). Anchoring bias was initially emphasized in a study by Tversky and Kahneman (1974), in which participants spun a wheel of fortune that generated a random number (0–100) and were then asked whether the percentage of African nations in the United Nations was higher or lower than that number (Strack et al., 2016). After making the comparative judgment, participants then estimated the numerical percentage (i.e., an absolute judgment). Surprisingly, those who obtained higher numbers from the wheel gave higher absolute estimates, showing that anchors can systematically bias judgments. Interestingly, the phenomenon is robust in the sense that it occurs even when the anchor is arbitrary or irrelevant to the task (Adaval and Wyer, 2011; Ariely et al., 2003) and when the domain expertise is high (Northcraft and Neale, 1987). In addition, multiple studies demonstrated the existence of anchoring bias, and its applicability extends to a wide range of domains, including IS (see, e.g., Wu and Cheng, 2011).

There are various explanations for the mechanisms behind the anchoring bias. One derives from Tversky and Kahneman (1974), who postulated that the phenomenon can be attributed to insufficient adjustment to an anchor. Specifically, they attributed the bias to the notion that decision-makers attempt to make a realistic estimate based on the anchoring value and adjust toward it but do not adjust sufficiently – in other words, they cognitively "move" only a short distance away from the anchor. The resulting estimate of the actual value (such as the absolute percentage of UN nations in Africa) then remains close to the anchor, even if the anchor is not a realistic starting point. Subsequent research has questioned the notion that insufficient adjustment can fully explain anchoring bias. More specifically, critics argue that if the anchoring effect arises solely from insufficient adjustment, then the anchor should influence only the absolute estimate, not the comparative judgment task (e.g., whether the proportion of African nations in the United Nations is higher or lower than 65%) that precedes it (Strack et al., 2016). In fact, there are indications that anchors often influence decisions on comparative questions even before any adjustment takes place (Jacowitz and Kahneman, 1995). If the comparative judgment is already influenced by the anchor, then the effect could not be explained solely through an insufficient adjustment in the absolute estimate (Strack et al., 2016).

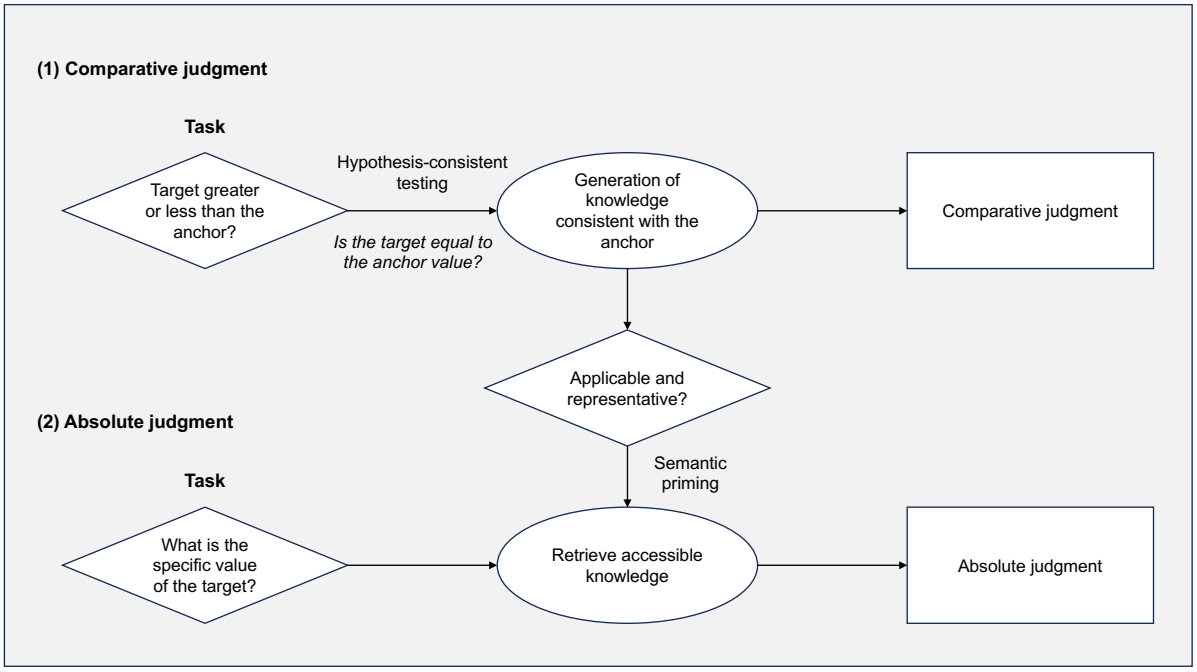


Figure 3: Selective accessibility model (adapted from Mussweiler and Strack, 1999; Strack et al., 2016)

Another explanation is the selective accessibility model (Figure 3) outlined in Mussweiler and Strack (1999). The selective accessibility model attributes the anchoring effect to a selective increase in the cognitive accessibility of anchor-consistent knowledge with respect to the target (Mussweiler et al., 2000). In the (1) comparative judgment task (i.e., "Target greater or less than the anchor?"), individuals perform hypothesis-consistent testing by evaluating whether the target value could match the anchor value (Mussweiler and Strack, 1999). In this process, humans seem to confirm anchor-consistent information more likely than anchor-inconsistent information (Klayman and Ha, 1987). This could involve, in the context of the experiment of Tversky and Kahneman (1974), thoughts like "I can easily recall several African nations that are United Nations members" for a high anchor (Strack et al., 2016). The generation of knowledge consistent with the anchor is thereby selectively increased as a result of this testing process (Mussweiler and Strack, 1999). Even if this testing process is unsuccessful and the hypothesis is ultimately rejected, it generates a semantic priming effect that facilitates access to consistent information for a subsequent task (Strack et al., 2016). Finally, the model postulates that, for the (2) absolute judgment task (i.e., "What is the specific value of the target?"), individuals assess the applicability and representativeness of knowledge with respect to the target and retrieve this easily accessible knowledge during target value estimation (Mussweiler and Strack, 1999).

Interestingly, Mussweiler et al. (2000) argued that providing knowledge inconsistent with the anchor may mitigate this effect. Specifically, they contended that the absolute judgment is

biased due to the accessibility of anchor-consistent knowledge. Activating anchor-inconsistent knowledge (e.g., additional disconfirming information) might therefore serve as a corrective strategy that could mitigate anchoring bias (Mussweiler et al., 2000).

The anchoring bias is particularly relevant for Paper VI, which focuses on activating such inconsistent knowledge and, more specifically, the potential of XAI to overcome anchoring bias. In particular, the paper investigates the interplay between explanations in XAI-based decision support and positioned anchors in a price estimation task.

2.5.2 Self-regulated learning theory

Humans constantly regulate their behavior across a wide range of domains, whether it is managing what they eat, controlling emotional responses, or making responsible decisions in risky situations. The processes by which the human mind exercises control over its own states, functions, and internal mechanisms are referred to as self-regulation (Baumeister and Vohs, 2004). Self-regulation involves the regulation of one's own thoughts, emotions, impulses, and task performances. Consequently, it is relevant for almost all forms of behavior in various domains (Baumeister and Vohs, 2004). Related research has, for example, explored how individuals regulate their eating habits (Johnson et al., 2012) as well as how self-regulation processes relate to sexual risk-taking behavior (Song and Qian, 2020).

One domain in which self-regulation appears to play a decisive role in success is higher education (Richardson et al., 2012). Compared to secondary education, in higher education, students usually have more freedom in organizing their learning, such as choosing courses of interest and creating their own schedules. While this flexibility offers advantages (e.g., aligning learning sessions with individual preferences), it also introduces challenges for students' self-regulation – more specifically, they must rely more on their own self-regulation skills. For example, students must take more responsibility for planning their learning efforts and monitoring and evaluating their learning progress (Vosniadou, 2020). This also applies to digital learning environments (Kizilcec et al., 2017), which offer greater flexibility in terms of time and location. At the same time, digital learning environments might pose specific challenges for learners, such as how learners manage their motivation and engagement without immediate support from instructors.

The research stream focusing on understanding the factors that influence individuals' learning processes is known as *self-regulated learning (SRL)* (see, e.g., Credé and Phillips, 2011; Pintrich, 2004; Wild and Schiefele, 1994). Within SRL, learners differ in their cognitive, metacognitive, and resource management skills (Panadero, 2017). These skills are in turn related to SRL strategies (Credé and Phillips, 2011; Pintrich, 2000). In this dissertation, the terms *skills* and *strategies* are used in the SRL context as follows: *skills* refer to the ability of the learner to self-regulate, while *strategies* are the actions the learner implements based on SRL skills. For example, students with well-developed metacognitive skills could more often activate prior

knowledge when planning learning activities (Pintrich, 2000) and therefore can identify gaps in content knowledge. In contrast, students with weaker metacognitive skills might be less able to identify these gaps. Likewise, students with strong time management skills could more often use regulation strategies and plan their learning (e.g., by creating and following structured study plans) (Pintrich, 2004), while poorer organization (i.e., structuring and planning) is associated with self-regulation failures such as procrastination (Steel, 2007).

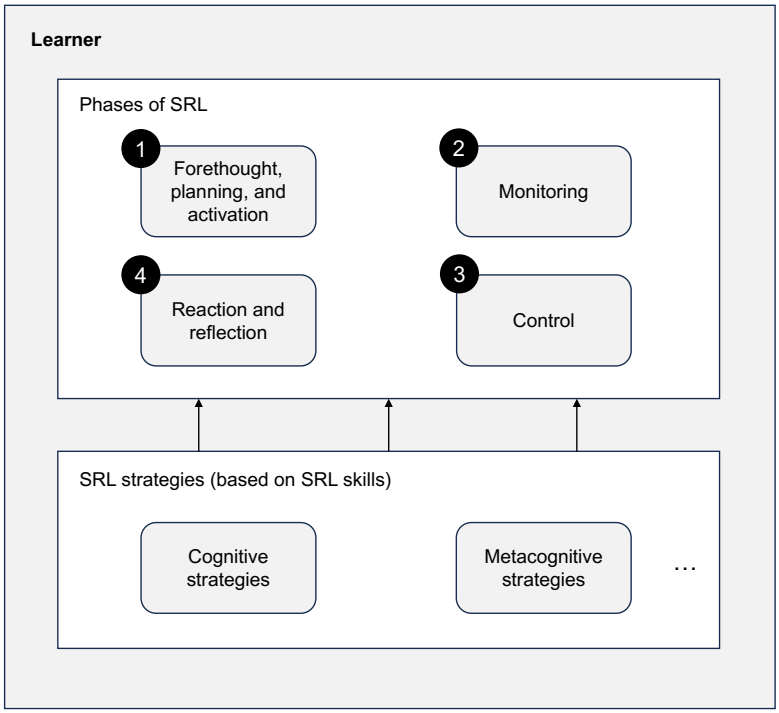


Figure 4: SRL phases based on Pintrich (2004) and their relation to SRL strategies (adapted from Paper IX)

The SRL literature has put forth multiple theoretical models describing how learners engage in self-regulation (Panadero, 2017). Two models particularly relevant to this dissertation are those of Zimmerman and Moylan (2009) and Pintrich (2004), which are similar in structure but different in focus. Zimmerman and Moylan's (2009) model is grounded in social cognitive theory and conceptualizes SRL as a cyclical process comprising three phases: (i) forethought, (ii) performance, and (iii) self-reflection. In contrast, Pintrich's (2004) model places a stronger focus on motivational aspects. The inclusion of both models in this work is motivated by their complementarity: Zimmerman and Moylan's (2009) broader structure aligns well with this dissertation's approach to DSS instantiation (Paper VII), while Pintrich's (2004) emphasis on motivation provides essential theoretical grounding for the operationalization of SRL strategies using the Motivated Strategies for Learning Questionnaire (MSLQ). The MSLQ is an important instrument presented in Pintrich et al. (1991) and used in Paper IX to measure the use of students' motivational orientations and the use of learning strategies in higher education.

Figure 4 outlines the SRL phases of the second relevant model based on Pintrich (2004) and also displays their relations to learners' SRL strategies. In the first phase of the model (*1: Forethought, planning, and activation*), learners plan their learning activities and activate their perceptions and prior knowledge related to the task. In the second phase (*2: Monitoring*), learners monitor themselves by reviewing their learning plans. Specifically, these monitoring processes involve metacognitive awareness of the self (e.g., self-observation of behavior by evaluating the effort and time invested in the task), the task itself, or the learning environment. This phase can also include learners recognizing that they may need help with the task (Pintrich, 2004). In the third phase (*3: Control*), learners regulate themselves. In this phase, learners select and adapt cognitive strategies, which may be reflected in increased or decreased effort and active help-seeking behavior (Pintrich, 2004). The fourth phase (*4: Reaction and reflection*) involves cognitive judgments about the progress and the effectiveness of the selected learning strategies (e.g., task planning or repetition of learning material). The effective use of these strategies thereby seems to depend on the skills of the learners. For example, learners with low levels of metacognitive skills may struggle to organize their learning during the forethought, planning, and activation phase (i.e., phase 1), while those with higher levels are more likely to activate prior knowledge, identify gaps in understanding, and plan how to address them (Pintrich, 2004). Metacognitive skills are therefore crucial to learning success and have a direct impact on learning outcomes (Pintrich, 2004). One of the general assumptions in SRL is that self-regulation activities mediate the relationship between learner or contextual characteristics (e.g., SRL skills and the classroom environment) and respective outcomes (Pintrich, 2004) as well as SRL failures, such as procrastination (Steel, 2007; Wolters, 2003).

In particular, procrastination is a phenomenon frequently mentioned in literature (see, e.g., Senécal et al., 1995; Wäschle et al., 2014; Wolters, 2003) and closely related to SRL as it reflects a form of failure in self-regulation (Tuckman, 2005). Academic procrastination refers to the postponement of study-related activities that one should but fails to complete within a desired time frame (Senécal et al., 1995). A defining characteristic of procrastination is the intentional delay of task-related activities, which often leads to last-minute efforts to meet deadlines (Schraw et al., 2007; Wolters, 2003). Consequently, procrastination can have severe consequences for academic performance (Steel, 2007; Tuckman, 2005; Wolters, 2003). Similar to the variability observed among students with respect to SRL, there are considerable individual differences in procrastination tendencies: while some students manage to complete academic tasks on time, others frequently procrastinate, potentially facing severe consequences (Tuckman, 2005). In general, students who frequently procrastinate stand in stark contrast to those considered self-regulated learners (Wolters, 2003).

SRL theory and its associated learning outcomes are relevant to Papers VII, VIII, and IX and play a crucial role in (a) instantiating an XAI-based DSS to support learners in digital learning environments and (b) evaluating the effects of such a DSS in the field.

3 Methodology

This section describes the overall methodological approach of this dissertation. Table 2 provides an overview of the main methods used across the included papers and identifies each paper's primary approach to validation and evaluation following the taxonomy of Doshi-Velez and Kim (2017). Unlike papers that employ predictive analytics and experimental research, Paper I and Paper II report the results of a meta-analysis that provides a comprehensive synthesis of the effect of XAI-based decision support on human task performance across multiple empirical studies in the literature. As both papers primarily include studies within controlled environments, they mainly focus on human-grounded evaluation (Doshi-Velez and Kim, 2017).

Table 2: Overview of methodologies used in each paper of this dissertation

Chapter of this dissertation	**Paper**	**Main research method**	**Main evaluation/ validation approach**
Chapter 1: Synthesizing the overall effect of XAI-based decision support on task performance	I	Meta-analysis	Human-grounded
	II	Meta-analysis	Human-grounded
Chapter 2: Validating and evaluating explanations in XAI-based decision support	III	Predictive analytics	Functionally grounded
	IV	Predictive analytics	Functionally grounded
	V	Online experiment	Human-grounded
	VI	Online experiment	Human-grounded
Chapter 3: Instantiating an XAI-based decision support system and evaluating its effects in the field	VII	Predictive analytics	Functionally grounded
	VIII	Field experiment (concept)	Application-grounded
	IX	Field experiment	Application-grounded

3.1 Predictive analytics

According to Shmueli and Koppius (2011), predictive analytics involves the use of statistical models and techniques such as data mining algorithms to generate empirical predictions and evaluate their quality. The use of predictive analytics in IS is manifold, including the comparison of competing theories, the synthesis of new theories, and the assessment of a phenomenon's predictability (Shmueli and Koppius, 2011). Papers III, IV, and VII use predictive analytics to assess the predictability of empirical phenomena. More specifically, Paper VII uses predictive analytics for a design-oriented research approach (see Baskerville et al., 2015; Hevner et al., 2004; Peffers et al., 2007 for more details on design science research). Papers V, VI, VIII, and IX, in contrast, make use of predictive analytics to build the basis for the investigations of human-XAI collaborations and resulting outcomes in XAI-based DSS.

The predictive analytics approach of these papers follows three consecutive steps in supervised ML based on Kühl et al. (2021). The first step is model initiation, which includes data acquisition to collect measurable and potentially relevant data for prediction. The data include public and utility-provided energy datasets (Paper III, Paper IV, Paper V), a publicly available automotive platform dataset (Paper VI), and a custom-built solution that tracks student characteristics, behaviors, and learning outcomes in a digital learning environment (Paper VII, Paper VIII, Paper IX). This step also involves data preprocessing and exploration, and the subsequent generation of features to predict the target. The second step involves the training of models to learn the relationship between independent variables (i.e., the features) and a dependent variable (i.e., the target to be predicted). The prediction tasks comprise future cross-buying behavior of existing customers (Paper III), household electricity consumption (Paper IV), students' exam performance (Paper VII, Paper VIII, Paper IX), household characteristics (Paper V), and used car market prices (Paper VI). The second step also involves evaluating the model's predictive power (i.e., performance estimation) using techniques such as cross-validation and out-of-sample testing and predictive performance metrics (see Hastie et al., 2009 for more details). The third step, if relevant, includes model deployment, which involves the actual integration of a model (e.g., through a webservice, as outlined in Paper VII) in a system. In summary, these steps served as the foundation for validating and evaluating XAI-based explanations and the instantiation of an XAI-based DSS.

Explanatory modeling, in contrast, traditionally employs statistical methods (most commonly regression) to test causal hypotheses that aim to explain how and why specific phenomena occur (Shmueli and Koppius, 2011). Unlike in predictive analytics, the resulting models must be interpretable to examine the relationships between variables (Shmueli, 2010). In the context of this dissertation, the papers employ explanatory modeling to (i) estimate whether a treatment is the cause of a change in outcomes and (ii) investigate potential moderators.

Contrary to what the term XAI might suggest, XAI relates to the concept of predictive modeling – or more broadly, predictive analytics – as it aims to explain a predicted outcome (Shmueli et al., 2025). As highlighted by Shmueli et al. (2025) and Rudin (2019), it is important to exercise caution when interpreting the outputs of XAI. First, when using XAI outputs in a sense of explanatory modeling, one must recognize that XAI explanations rely solely on correlations. As a result, they can become disconnected from reality and yield explanations or suggest changes that do not actually affect real-world outcomes. In other words, there is no guarantee that these explanations reflect causal relationships (Shmueli et al., 2025). Second, the outputs of post-hoc applied XAI methods are only approximations of underlying ML-based prediction models, and they may not always accurately reflect the ML model being explained, as highlighted by Rudin (2019).

3.2 Experimental research

This dissertation uses experimental research for the Papers V, VI, VIII, and IX, which focus on the human-centric evaluation of explanations and the real-world effects of XAI-based DSS. These papers utilize two main evaluation approaches following Doshi-Velez and Kim (2017): (i) application-grounded evaluations, conducted through a field experiment in a digital learning environment involving students (Paper VIII and Paper IX); and (ii) human-grounded evaluations, implemented in a more controlled environment (i.e., online experiments) to investigate mechanisms for explanations in XAI-based decision support (Paper V and Paper VI).

An experiment is a quantitative research method used to examine cause-effect relationships (Recker, 2021). To this end, experiments vary a presumed cause and observe whether changes in this cause lead to changes in the effect (Shadish et al., 2002). This variation is introduced through a *treatment*, an experimental stimulus given to the treatment group but not to the control group (Recker, 2021). In this dissertation, the treatments involve (i) AI-based decision support implemented as ML-based predictions and (ii) XAI-based decision support, presented to users either through XAI visualizations (Papers V and VI) or embedded in an XAI-based DSS that provides digital feedback (Paper VIII and Paper IX).

To minimize the influence of confounding factors that could offer alternative explanations for differences in outcomes between groups, experimental research offers certain techniques. One technique is *random assignment* to groups, which helps to ensure that the groups are, on average, similar to each other in terms of preexisting conditions (Recker, 2021). Specifically, there should be no statistically significant differences in pre-existing conditions between the groups before treatment exposure. Therefore, if the random assignment works as intended, any differences observed in outcomes between groups are likely due to the treatment (Shadish et al., 2002). To further strengthen the argument that differences in outcomes are attributable to the treatment rather than to extraneous factors, one can employ *controls*, such as by collecting additional variables to better isolate the effects of the treatment (Recker, 2021). Consequently, all papers employing experimental research randomly assigned participants to control and treatment groups to reduce the risk of alternative explanations for observed differences between groups. Additionally, they collected variables to control for potential confounding factors (e.g., *age* in Paper V or *domain literacy* in Paper VI).

Experiments can be implemented in various *experimental setups*. Laboratory experiments typically provide a high level of control over external influences (e.g., environmental conditions), which increases internal validity (Cook and Campbell, 1979), meaning the validity of inferences about whether there is a causal link between the treatment (i.e., the independent variable) and the outcome (i.e., the dependent variable) (Coolican, 2009; Shadish et al., 2002). However, this high level of control often comes at the expense of less external validity, specifically ecological validity (i.e., the generalizability to other setups), as laboratory experiments are usually less reflective of a population's real-world conditions (Bryman, 2012; Coolican, 2009). Online

experiments (i.e., participants solve artificial tasks designed for experimental purposes in a digital environment) also allow partial control of external factors as the task environment can be controlled (Fink, 2022). Furthermore, compared to a physical laboratory environment, participants in online experiments can be recruited more efficiently using online labor platforms (Fink, 2022). Field experiments, in contrast, take place in a natural environment and aim to obtain the natural behavior of the population studied (Coolican, 2009; Fink, 2022). This could involve solving a quiz or watching a video in a digital learning environment of a real-world university course. However, such a setup often yields a comparatively reduced control over extraneous factors (e.g., participants in the control group behave differently as they know about the treatment and try to emulate it), posing threats to internal validity (Coolican, 2009). In this dissertation, the experimental setups chosen aimed to balance control over extraneous factors and real-world conditions, depending on the specific research focus of each study. Investigating cognitive influencing factors – such as user comprehensibility and preferences (Paper V) and biases and heuristics (Paper VI) – required a more controlled environment to better isolate potential extraneous factors. In contrast, the primary objective of the study on XAI-based feedback (field study concept in Paper VIII and the report of the experiment implemented in Paper IX) is to understand the effects of XAI-based feedback on learner behavior and outcomes under real-world conditions. Consequently, Paper V and Paper VI used online experiments, while Paper VIII and Paper IX used a field experiment as the experimental setup.

In addition to varying the experimental setups, the papers that rely on experimental research implemented different *experimental designs*. One such design, described in Paper VIII and applied in Paper IX, is the difference-in-differences (DID) design. The DID design can be used to isolate the effect of a treatment with respect to changes in outcomes over time (Angrist and Pischke, 2009). In the most basic setup, the control and treatment groups are observed in two time phases, namely a pre-intervention and post-intervention phase, but only the treatment group receives the treatment in the post-intervention phase (Imbens and Wooldridge, 2009). To isolate the treatment effect, the design subtracts the average change in the outcome (i.e., the response trend) in the control group between the two phases from the average change over time in the treatment group (see Figure 5, adapted from Angrist and Pischke, 2009). This double differencing (i.e., change in outcomes across phases for each group and how that change differs between groups) enables the DID approach to increase control over sources of bias in treatment effect estimation (Imbens and Wooldridge, 2009), enhancing internal validity. It does so in two ways: first by accounting for baseline differences between groups prior to treatment exposure, and second by controlling for extraneous factors associated with time trends, through the inclusion of the control group that is observed in both phases.

To make this more concrete, consider the example of online learning adapted from Paper VIII and Paper IX. Among other effects, the field experiment investigated whether personalized feedback increases the time students spend on a learning platform. The treatment group received personalized feedback, the control group did not, and the average of learning time is

estimated in the pre-intervention and post-intervention phase. If both groups differ in learning time before a treatment is provided (i.e., baseline differences) and increase their learning time in the post-intervention phase due to an external factor, e.g., because an exam is approaching, the DID design still allows control over such factors. It does so by comparing the extent to which the treatment group altered learning time relative to the control group while also (i) isolating baseline differences in learning time in the pre-intervention phase, and (ii) controlling for extraneous factors, such as upcoming deadlines, as they equally affect both groups.

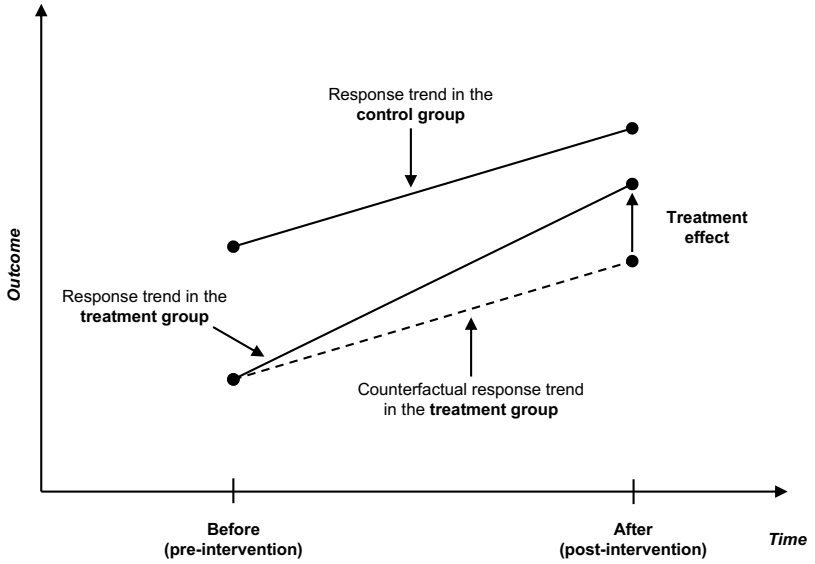


Figure 5: DID design (based on Angrist and Pischke, 2009)

With respect to data analysis, the main technique used for human- and application-grounded evaluations is ordinary least squares (OLS). OLS estimates the coefficients of a linear regression model by minimizing the sum of squared differences between the fitted and actual observed values (i.e., the residuals) of the continuous dependent variable (Wooldridge, 2016). In a DID design such as the one described above, individuals are observed at multiple points in time, which is why it yields panel data (Recker, 2021). In this setting, an OLS regression model specification can be written as follows (adapted from Paper IX):

$$y_{it} = \alpha_i + IN_{it}^{PostPhase} \times (\beta_0 + \beta_1 T_i^{XAIFeedback}) + \varepsilon_{it} \quad (1)$$

where y_{it} denotes the time spent online for an individual i at time t (i.e., week). This model incorporates the fixed effects parameter α_i for each individual to control for fixed unobserved differences between individuals (see, e.g., Wooldridge, 2016, for more details). The dummy variable $IN_{it}^{PostPhase}$ is coded as 0 during the pre-intervention phase and 1 during the post-intervention phase (see Figure 5 for the time periods of the DID design). $T_i^{XAIFeedback}$ is also a dummy variable that indicates whether an individual i is in the control group ($T_i^{XAIFeedback} = 0$)

or in the treatment group ($T_i^{XAIFeedback} = 1$). Finally, ε_{it} denotes the error term. As the data involve repeated measures per individual, the analysis in Paper IX employs clustering of standard errors at the individual level to mitigate potential biases in determining statistical significance (Croissant and Millo, 2019).

Another applied experimental design used in this dissertation is the factorial design, which is characterized by the inclusion of multiple treatments (i.e., two or more independent variables). An independent variable describes a factor, while all values of interest in this variable denote factor levels (Recker, 2021). Figure 6 shows an exemplary 2×2 factorial design based on Paper VI that a C2C retail platform could use to investigate how price tags acting as anchors and AI-based price predictions for a fair price influence customers' price estimates for a clothing item. This design would involve two independent variables: *AI-based decision support* and *anchoring*, each with two factor levels (i.e., absence versus presence), resulting in four experimental conditions. This setup could help a platform operator not only to understand customers' willingness to pay and the influence of price anchors, but also to assess how AI-based price predictions affect customer estimates and whether such support can mitigate the anchoring effect. Given that a factorial design allows the testing of multiple independent variables and their interaction effects, Paper V and Paper VI (similar to the example outline above) use factorial designs to examine the interaction between multiple independent variables.

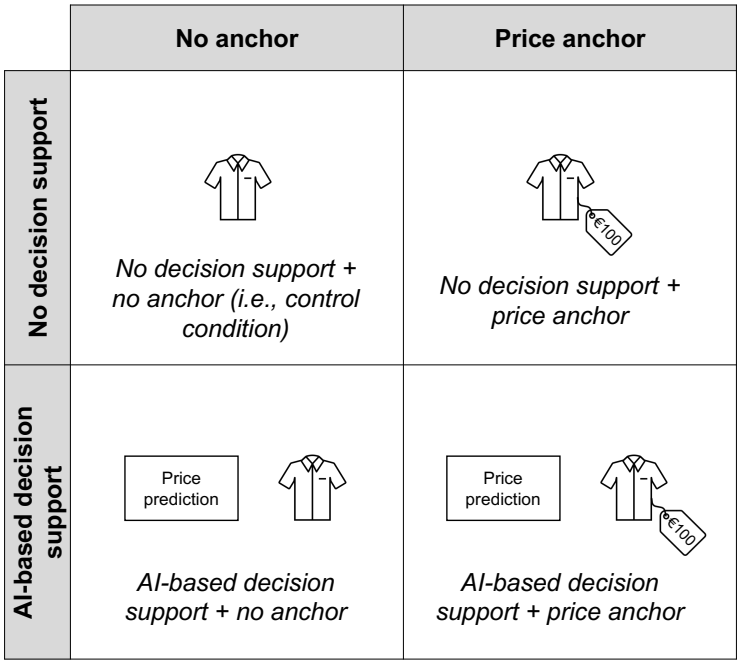


Figure 6: Exemplary 2×2 factorial design (based on Haag, Stingl, et al., 2023; Paper VI)

An experimental design like that in Figure 6 involves independent observations and provides cross-sectional data, which means that the behavior of individuals is observed at a single time point (Recker, 2021). The corresponding regression formula can be defined as follows:

$$y_i = \beta_0 + \beta_1 T_i^{DecisionSupport} + \beta_2 T_i^{Anchor} + \beta_3 (T_i^{DecisionSupport} \times T_i^{Anchor}) + \varepsilon_i \qquad (2)$$

where y_i is the dependent variable for the individual i and β_0 denotes the intercept, which is constant across all individuals in this case. The variables $T_i^{DecisionSupport}$ and T_i^{Anchor} represent the dummy variables indicating the presence (1) or absence (0) of decision support and a price anchor, respectively. The interaction term $(T_i^{DecisionSupport} \times T_i^{Anchor})$ captures how the effect of decision support on the outcome changes in the presence of the anchor (and vice versa). The error term ε_i accounts for the variance not explained by the model.

3.3 Meta-analysis

Paper I and Paper II conduct a meta-analysis to holistically examine the effect of XAI-based decision support on task performance, drawing on results from multiple empirical studies in the literature. In addition to estimating the overall effect of XAI-based decision support on task performance, both papers investigate the explanation type and the risk of bias in studies as potential moderators.

Meta-analysis is a quantitative research method used to combine results from multiple empirical studies (Higgins et al., 2024). Unlike qualitative reviews, which remain more prevalent in leading IS publication venues (Templier and Paré, 2018), meta-analysis relies on objective data, synthesizing results across studies to provide more precise effect estimates than individual studies (Higgins et al., 2024; King and He, 2005). Specifically, meta-analysis can aid in theory testing, help resolve conflicting findings, and identify moderators that explain variations in outcomes (King and He, 2005). This dissertation applies meta-analysis for the latter two reasons: (a) to examine the conflicting findings regarding the impact of XAI-based decision support systems on task performance and (b) to analyze potential moderators. Prior studies have reported substantial variability in effect sizes, which supports a systematic approach to clarifying these discrepancies (see Paper II for further details).

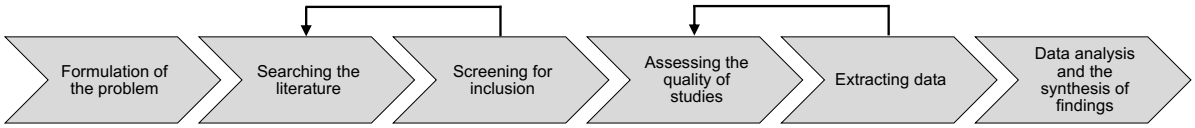


Figure 7: Steps in meta-analysis (steps adapted from Templier and Paré, 2018 who built on Cooper, 2009)

The procedure for the meta-analysis in this dissertation follows six steps, as outlined by Templier and Paré (2018) building on Cooper (2009), that are partly iterative (see Figure 7). First, the *formulation of the problem* involves defining the core focus of the investigation and determining which studies are relevant for inclusion. In Paper I and Paper II, this concerns user studies that report task performance measures and use explanation types that are particularly relevant for user XAI-based decision support (see Section 2.3.1). Second, the next step is *searching the literature* to identify potentially relevant studies. For the relevant papers in this dissertation, this includes an iterative search string development. Third, *screening for inclusion* involves applying criteria to (i) assess basic suitability – where the resulting set of studies served as the basis for a forward-backward search (as indicated by the backward arrow in Figure 7) – and (ii) further distinguish between suitable and unsuitable studies, such as by excluding those that do not report objective decision performance. Curating the study sample is especially important to reduce heterogeneity, as meta-analysis is deemed unsuitable if heterogeneity between studies is high (Paré and Kitsiou, 2017). Fourth, the process involves *assessing the quality of studies*, including external evaluations (e.g., verifying the consistent exclusion of studies with unsuitable designs). Fifth, *extracting data* from the selected studies involves collecting study characteristics, decision support attributes, and statistical results. Since standard deviations were not always reported or could be calculated from the given information, the authors were contacted by email to retrieve this information. Furthermore, this step involves categorizing studies by (a) the type of explanation and (b) the risk of bias in randomized trials using version 2 of the Cochrane risk-of-bias tool (RoB 2) (Sterne et al., 2019), and additional quality assessments of these categorizations through external evaluators (as indicated by the backward arrow in Figure 7). Sixth, *data analysis and the synthesis of findings* comprise estimating the pooled effect sizes, assessing heterogeneity across studies, and conducting moderation analyses using regression models to draw overarching conclusions.

Compared to the papers that employ experimental research, Paper I and Paper II employ random-effects models to incorporate the variability between studies examining the impact of XAI-based decision support on human task performance. Unlike fixed-effects models, which assume a single and common true effect size between studies, random-effects models consider that the true effect size may vary across studies due to differences in study design, sample characteristics, and tasks (Harrer et al., 2021; Higgins et al., 2024). To compare the different study conditions – such as decision support based on AI vs. XAI – the analysis computes standardized mean difference (SMD) using Hedges' g (Hedges, 1981). Hedges' g is a bias-adjusted effect size metric designed to correct for inflation in small-sample studies ($N \leq 20$). In addition, the models estimate an overall average effect on task performance, referred to as the "pooled effect" across all studies included in each analysis.

4 Main results

This section summarizes the main results presented in each chapter. It first presents the results from Paper I and Paper II, which conduct a meta-analysis to synthesize the overall effect of XAI-based decision support on human task performance. Specifically, they are presented within the scope of:

> **Chapter 1**: Synthesizing the overall effect of XAI-based decision support on task performance (addressing **Research Objective 1**)

Then, the section outlines the main results of Papers III–VI that conduct in-depth investigations into explanations in XAI-based decision support, focusing on method-centric validation and human-centric evaluation. These results are outlined within the scope of

> **Chapter 2**: Validating and evaluating explanations in XAI-based decision support (addressing **Research Objective 2**)

After this, the main results of Papers VII–IX are presented, which describe the instantiation of an XAI-based DSS that provides personalized feedback by addressing differences between learners. These results also include evaluation of the system's effects in an actual university course in a digital learning environment. These results are presented as part of

> **Chapter 3**: Instantiating an XAI-based decision support system and evaluating its effects in the field (addressing **Research Objective 3**)

4.1 Chapter 1: Synthesizing the overall effect of XAI-based decision support on task performance (Paper I* and Paper II**)

Explanations embedded in IS have the potential to enhance users' decision-making performance (Gregor and Benbasat, 1999). Despite this potential, the corpus of empirical studies on the utility of explanations in XAI-based decision support finds mixed effects. Indeed, some studies have observed considerable increases in task performance when humans team up with AI-based decision support accompanied by explanations (i.e., XAI-based decision support) (see, e.g., Bansal et al., 2021; Lai et al., 2020), while others report no or adverse effects (Bauer et al., 2023; Carton et al., 2020). A meta-analysis by Schemmer, Hemmer, et al. (2022) reported

∗ F. Haag (2025). "The Effect of Explainable AI-based Decision Support on Human Task Performance: A Meta-Analysis." *Proceedings of the 23rd Annual Pre-ICIS Workshop on HCI Research in MIS*. Bangkok, Thailand.

∗∗ F. Haag (2026). "How Explanations from XAI-based Decision Support Affect Human Task Performance: A Meta-Analysis." *Journal of Decision Systems* 35 (1). DOI: 10.1080/12460125.2026.2616693.

some initial results in the context of XAI-based decision support and identified an overall positive effect for human task performance, but this meta-analysis was limited to a set of nine articles. The extent to which such support affects human task performance across a more recent body of empirical studies remains unclear. In this context, several conditions may play a crucial role in the effectiveness of XAI-based decision support. One of these conditions, which has been repeatedly identified as important in related research, is the type of explanation (see, e.g., Herm, 2023b). In particular, *Why?*/*Why-not?*/*How?* explanations (represented through feature attribution and feature importance methods) as well as *How-to?* (represented through counterfactual methods) and *What-else?* (represented through example-based methods) explanations are frequently examined in the context of XAI-based decision support (see, e.g., Bansal et al., 2021; Bauer et al., 2023; Buçinca et al., 2020; Herm, 2023b; Mohseni et al., 2021; Wang and Yin, 2021). However, it is also unclear how the respective types of explanations affect human task performance across empirical studies.

In response to these shortcomings in the literature, Paper I and Paper II present the findings of a meta-analysis that investigates how XAI-based decision support affects human task performance. Because most studies assess task performance through tasks requiring participants to choose among two decision alternatives, the papers focus specifically on binary classification tasks. While Paper I reports the results of the analysis based on an initial search conducted in January 2023, Paper II updates this analysis with a search conducted in August 2024. Given that both papers share the same data collection and statistical analysis procedures, the following results focus on Paper II as the updated analysis contains the most recent findings.

The literature search resulted in 1,015 articles, identified from the databases Scopus, Web of Science, and EBSCOhost. Papers were filtered to meet the requirements of the meta-analysis and to ensure low heterogeneity between included studies (e.g., minimizing differences in study designs and study tasks). The final literature review of Paper II in 2024 yielded 33 behavioral studies drawn from 27 articles. Data were recorded separately for participants under three conditions: (i) without any decision support (denoted as no decision support), (ii) with support provided through predictions (denoted as AI-based decision support), and (iii) with predictions combined with explanations (denoted as XAI-based decision support). The results of the comparison between no decision support (control group results) and XAI-based decision support* (treatment group results) show an SMD of 0.523, indicating a moderate effect (Harrer et al., 2021) that significantly differs from zero (t value = 4.45, p value <0.001). In other words, this result indicates that AI-based decision support that also includes explanations can effectively support individuals in completing tasks when compared to participants who received no support. The second analysis concerned the comparison between AI-based decision support (control group results) and XAI-based decision support (treatment group results), which allowed for a better understanding of the marginal effect of explanations in AI-based

* All subsequent reports on statistical significance and effect sizes for Paper II correspond to outputs from random-effects regression models.

decision support. In this comparative setting, the result displays a small effect of 0.086 SMD (Harrer et al., 2021) through the XAI treatment that significantly differs from zero (t value = 2.39, p value = 0.020). Interestingly, given that the SMD is near zero, the results indicate that there is overall little to no effect of explanations in XAI-based decision support across the empirical studies included.

Beyond the main treatment effect, the statistical analysis also explored how the risk of bias in study results and the type of explanation influence the effect of explanations on task performance. Here, this comparison focuses on AI-based vs. XAI-based decision support to identify additional factors that may help explain the observed effect. Table 3 presents the results of the risk of bias analysis based on the "RoB 2" tool by Sterne et al. (2019), which categorizes the risk of bias in studies as "Low", "Some concerns", or "High". The analysis revealed a notable trend: human task performance seems to increase as the risk of bias in the studies increases. Studies assessed as having a "Low" risk of bias had a small, non-significant effect (SMD = 0.051, p_{type} value = 0.380). Studies categorized as "Some concerns" exhibited a slightly larger effect (SMD = 0.108, p_{type} value = 0.020). Most notably, studies with a "High" risk of bias report a substantial and significant increase in task performance (SMD = 0.281, p_{type} value = 0.002), indicating that studies with higher bias result in a higher treatment effect estimation. A test on subgroup differences confirmed that the risk of bias in studies significantly influences the effect estimate ($p_{subgroup}$ value = 0.025). This result therefore reinforces the notion that the actual impact of explanations in XAI-based decision support on task performance is small when viewed across studies. Another question of this analysis was whether the explanation type affects task performance differently. A corresponding analysis found no significant effect of any explanation type on task performance. The subgroup test also revealed no significant differences between explanation types ($p_{subgroup}$ value = 0.490).

Table 3: Risk of bias assessment for the comparison between AI-based vs. XAI-based decision support (adapted from Haag, 2026; Paper II)

Risk of bias	**Low**	**Some concerns**	**High**
SMD	0.051	0.108	0.281
95% CI	[–0.036; 0.137]	[–0.039; 0.254]	[0.111; 0.450]
p_{type} value	0.380	0.020	0.002
$p_{subgroup}$ value		0.025	

Notes: The table compares SMDs between AI and XAI-based decision support systems under different risk-of-bias levels. p_{type} denotes level-specific p-values, while $p_{subgroup}$ represents the test for differences between subgroups.

In sum, the results of Paper I and Paper II fulfill **Research Objective 1** by synthesizing the overall effect of XAI-based decision support on human task performance from the body of empirical studies in the literature. Specifically, the results show that the additional effect of explanations, compared to sole AI-based decision support, is small overall. Further analyses also reveal that (a) the treatment effect estimation increases as the risk of bias in study results increases and (b) the type of explanation cannot explain the variation in the overall effect across studies. However, these results should not be interpreted as evidence that explanations in XAI-based decision support are negligible or that the explanation type plays no role in differences in task performance as study results vary considerably. Rather, it motivates further investigations to better understand the mechanisms that shape the effectiveness of XAI-based decision support.

4.2 Chapter 2: Validating and evaluating explanations in XAI-based decision support (Papers III–VI)

The main findings of Chapter 2 focus on assessing explanations in XAI-based decision support by examining mechanisms for the effectiveness of XAI-based decision support. These mechanisms pertain to both method-centric validation of XAI explanations and factors related to human-centric evaluation, together addressing **Research Objective 2**.

4.2.1 Method-centric validation: An analysis of the robustness of XAI explanations (Paper III*)

Feature attribution methods can provide insights to users by explaining how the input features contributed to the output of a particular model (see, e.g., Bouazizi and Ltifi, 2024). Whether such methods explain the results of ML outputs well from a method-centric point of view has frequently been analyzed within the scope of functionally grounded assessments using quantitative measures (Schwalbe and Finzel, 2023). However, many of these studies focus solely on XAI method properties (e.g., *stability*) for classification tasks (see, e.g., Velmurugan et al., 2021). Although insightful, these assessments should be extended to include a test on explanation robustness using real-world data to complement prevailing approaches. Such an assessment could not only help to assess the quality of explanations but also test whether explanations hold for future points in time, thereby testing their validity in real-world settings.

Paper III outlines such a robustness validation of feature-attribution-based explanations for the use case of cross-buying prediction in energy retailing. To this end, the study used a dataset from a large European utility company comprising N = 220,185 active customers in

∗ F. Haag, K. Hopf, P. Menelau Vasconcelos, and T. Staake (2022). "Augmented Cross-Selling Through Explainable AI—A Case From Energy Retailing." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

the years 2012–2017, covering 547,848 utility service contracts. Using these data, the study extracted features (e.g., power consumption, revenue for a specific year and geographic features such as the distance to businesses); trained several ML approaches on data from a given year; and finally predicted whether existing customers with a power, Internet, or cable TV contract would buy a new Internet or TV contract in the following year (i.e., cross-purchases). The dataset exhibited a highly imbalanced class structure, ranging from 0.5% to 1.9% of customers cross-purchasing new contracts each year (encoded in the positive class for buying customers and the negative class for non-buying customers). To address this class structure, the study applied the balanced random forest (Chen et al., 2004) and the random under-sampling boost (Seiffert et al., 2010) classifiers, both of which use random undersampling to balance the dataset. Finally, the XAI method SHAP explained for the case "Existing power customer buys a new TV contract" how each feature contributed to the respective predicted class. Specifically, the explanation relied on the balanced random forest model that was trained on data from 2016 to predict cross-purchases of TV contracts in 2017.

The subsequent validation comprised two analyses. The first is a baseline assessment and tested explanation robustness similar to the *stability* metric by examining whether SHAP produced internally consistent feature attributions across 10 random subsamples for observations of the same class (i.e., the positive class representing actual cross-purchasers). This analysis has the assumption that a robust explanation for a feature should not considerably differ for buyers across different data subsets. To this end, the analysis examined differences in SHAP values across test folds for customers who made cross-purchases (i.e., true positives and false negatives). Feature attributions were considered robust if (i) they showed no significant differences for buyers or (ii) any observed differences were statistically significant but had only a small effect size. The results of the analysis showed that among the top 20 most influential features there are no significant differences; if differences were significant, this was only with small effect sizes across subsamples. The second analysis examined whether the detected relationships persisted in future years. For this purpose, the study formulated assumptions based on the top 10 SHAP-assigned feature attributions and their impact on model output (e.g., customers who signed a TV contract in 2017 had significantly higher revenue in the previous year compared to those who did not). It is important to note that these assumptions reflect real-world hypotheses derived from model explanations, which may – but do not necessarily – reflect causal relationships in cross-buying behavior. To test whether these assumptions held for the following year, the study conducted statistical tests on differences between buyers and non-buyers for the respective features in the target year (i.e., 2017 in this case). The results show that for 8 out of 10 features, there was a significant difference between buyers and non-buyers, indicating that the relationships identified through SHAP-based explanations were largely persistent for the subsequent year.

> In summary, the results of Paper III demonstrate that the XAI method SHAP can provide robust explanations for the use case of cross-buying prediction in the energy retailing domain. Specifically, the study tested feature attribution explanation robustness by (a) assessing the similarity between subsets of similar observations (i.e., based on properties for explanation quality) and (b) the temporal persistence of patterns detected by the XAI method SHAP. Ultimately, these results suggest that there are XAI methods (in this case SHAP) that can provide robust explanations whose validity persists over a specified time frame.

4.2.2 Method-centric validation: A property-based validation approach for XAI explanation quality for various explainer methods (Paper IV*)

The validation of XAI explanations based on method properties is an integral part of XAI research. Such validations offer a systematic approach to formally examining explanation quality to determine whether explanations are *consistent*, *stable*, or *faithful* (i.e., validating properties of explanation quality) (Schwalbe and Finzel, 2023). Although numerical prediction is a common task and various explanation methods exist, many validation approaches focus solely on (a) classification tasks (see, e.g., Alvarez Melis and Jaakkola, 2018) and (b) specific types of explanation methods (e.g., post-hoc methods; see Schlegel et al., 2019).

To alleviate these limitations, Paper IV presents an approach that aims to extend prevailing validation methods, with a specific focus on numerical prediction tasks and method-agnostic assessments of feature attribution methods. The paper draws on established conceptual definitions of metrics from the literature, specifically (i) *model fit*, (ii) *consistency*, (iii) *stability*, and (iv) *faithfulness*, and implemented corresponding computational rules. For the first metric, *model fit*, the approach utilizes well-known numerical prediction quality measures. Although it is distinct from well-known explanation quality metrics (Robnik-Šikonja and Bohanec, 2018), this metric is included as related work suggests that a well-performing model is more likely to generate explanations that align with domain concepts because it reflects at least some concepts behind the domain (Štrumbelj et al., 2009). For the second metric *consistency*, measurement for non-post-hoc methods can be challenging, but also when the models use different features (Molnar, 2025). Hence, deviating from the conceptual definition, the approach proposes calculating *internal consistency*, which evaluates consistency across uniformly distributed subsamples. The third metric, *stability*, draws on the conceptual definition of Robnik-Šikonja and Bohanec (2018) in that the approach implements the metric by clustering similar observations and subsequently measuring differences in their explanations. It is important to note that *internal consistency* and *stability* are therefore similar in the context of this paper but different in calculation. The fourth metric, *faithfulness*, is measured through perturbation analyses, which examine changes in predictions in response to input variations. *Faithfulness*

* F. Haag, K. Hopf, and T. Staake (2026). "Validating Explainer Methods: A Functionally Grounded Approach for Numerical Forecasting." *Journal of Forecasting* 45 (2), pp. 819–836. DOI: 10.1002/for.70060.

(or fidelity) is most often evaluated in the context of post-hoc explanations of black-box models, where it measures how well the explanation reflects the underlying model's behavior (Molnar, 2025). To ensure completeness and comparability across model classes, *faithfulness* was also assessed for inherently interpretable models. In this setting, such models are not the primary target of *faithfulness* checks but instead serve as a baseline check, given that their explanations could be considered *faithful* by design (Miró-Nicolau et al., 2025; Rudin, 2019). Finally, a trade-off calculation between these metrics based on the geometric mean enables a comparative assessment of explainer methods through a single performance measure.

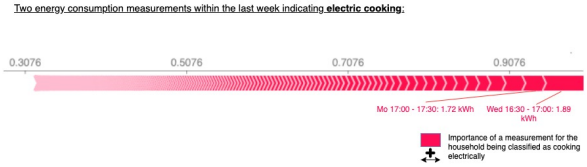
To demonstrate the applicability of the approach, the paper relied on the use case of electricity demand prediction in the context of energy benchmarking. This demonstration assessed explanations of methods with different characteristics, namely explainable boosting machine (EBM), multiple linear regression, LIME, and SHAP. The results showed that, for a given set of property weightings, SHAP and EBM achieved the highest overall performance among the metrics, followed by LIME, while multiple linear regression ranked lowest. Interestingly, EBM is consistently highly ranked across all metrics and provided the most *stable* explanations, while also being an intrinsically interpretable method.

> In sum, Paper IV introduced a property-based validation approach for feature attribution explanations specifically tailored to numerical prediction tasks. The results demonstrate that the XAI method SHAP and the interpretable ML approach EBM provide explanations that perform particularly well with respect to well-known properties from the literature. Overall, these results indicate that state-of-the-art explainer methods can provide valid explanations.

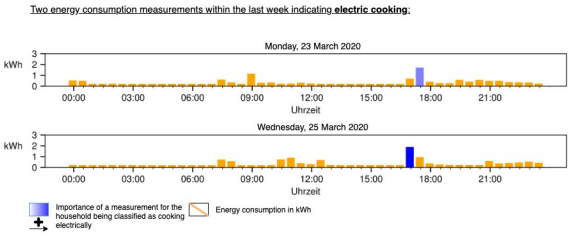
4.2.3 Human-centric evaluation: An assessment of user comprehensibility and preferences in XAI visualizations (Paper V*)

XAI-based decision support is frequently provided through ML-based predictions accompanied by XAI visualizations (see, e.g., Herm, 2023b). These visualizations help users interpret the model's decisions and derive new insights related to the task at hand, such as by presenting explanations in the form of feature attributions (see, e.g., Lai et al., 2020). Given their role in bridging the gap between the explanation provided by an XAI method and human cognitive processes (Fürnkranz et al., 2020), the way in which attributions are visualized may be an important factor for the effectiveness of XAI-based decision support. Specifically, understanding how different visualizations influence user comprehensibility, preferences, and decision-making could contribute to developing more effective and human-centric support.

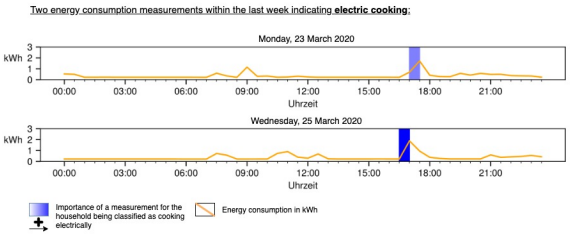
∗ J. Wastensteiner, T. M. Weiss, F. Haag, and K. Hopf (2021). "Explainable AI for Tailored Electricity Consumption Feedback – An Experimental Evaluation of Visualizations." *Proceedings of the 29th European Conference on Information Systems*. Marrakech, Morocco (virtual).



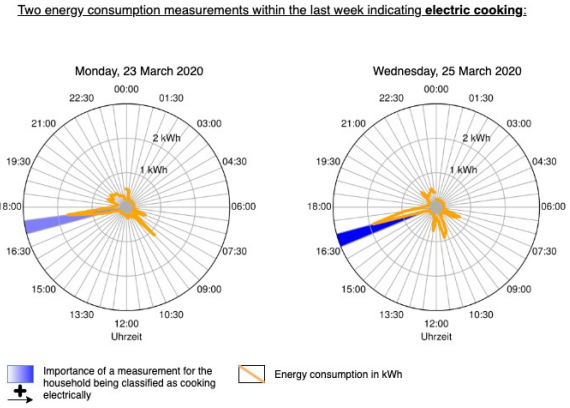
(a) SHAP force plot



(b) Bar diagram



(c) Line diagram



(d) Polar diagram

Your electricity consumption from 23 March 2020 to 29 March 2020 indicates that your household **cooks electrically**. For example, your electricity consumption during the following periods of that week (sorted by importance) indicates this:

- 1.89 kilowatt hours (kWh) on Wednesday from 16:30 to 17:00
- 1.72 kilowatt hours (kWh) on Wednesday from 17:00 to 17:30

(e) Generated text

Figure 8: XAI visualizations (adapted from Wastensteiner et al., 2021; Paper V)

Paper V reports the findings from an online experiment with N = 152 participants that examined the user comprehensibility and preferences of XAI visualizations for electricity consumption feedback. The experiment was preceded by a design phase that aimed to derive requirements for XAI and electricity consumer feedback visualizations from related literature. This design phase led to the development of five XAI visualizations, explaining ML models that predict household characteristics (Figure 8). Specifically, the ML models predicted whether a household uses electric cooking, has occupants at home on typical days, or heats water with electricity. Then, the XAI method SHAP by Lundberg and Lee (2017) generated the content for the XAI visualizations by deriving electricity consumption time intervals that could be relevant for the ML models predicting the respective class (i.e., feature attributions). The paper thereby refers to all visualizations as XAI visualizations because they are based on values derived from SHAP. Finally, the user study evaluated the visualizations in two phases. The first phase assessed comprehensibility through subjective understanding, self-reported mental effort, and the reading and answer times. Task performance was measured through objective understanding (i.e., in terms of the number of correct answers) and within the scope of the paper subsumed under comprehensibility. In four memorization tasks, participants first interpreted one of the five XAI visualizations and then subsequently classified statements without the visualizations on (i) electricity consumption at specific times, (ii) the ML prediction, and (iii) the explanation as correct or incorrect. The second phase focused on identifying user preferences for the visualizations through a conjoint experiment, whereby participants had to decide upon a preferred feedback visualization. The stimuli varied between the existence of explanatory text, two variants of energy-saving tips, and the presence of a chatbot frame. Out of a total of 40 possible stimuli, each participant received five randomly selected choice sets, with each set containing two visualization variants to choose from.

The results in Paper V of the first phase assessing the comprehensibility of XAI visualizations using memorization tasks indicate that there is overall little difference between the visualizations with respect to reading and answer times. However, an OLS multiple linear regression model* with the SHAP force plot (Figure 8a) as the reference level found that all other visualizations led to a higher task performance. This difference in task performance was statistically significant for the line diagram (19.77% more correct answers, p value = 0.003). This finding was also in line with the results on subjective comprehensibility: participants assigned better grades (from 1 = best to 6 = insufficient) in terms of understandability for all XAI-based visualizations when compared to the standard SHAP force plot. This difference was statistically significant for the line diagram (10.98% lower grades, which indicates better understanding, p value = 0.017) and text-based visualization (10.68% lower grades, p value = 0.012). The conjoint experiment in the second phase examining user preferences found that participants preferred the line and bar visualizations most. Specifically, a logistic regression using maximum-likelihood estimation

* The reported p values and effect sizes in percentages in this paragraph correspond to the estimated regression coefficients rather than a direct mean comparison between groups.

and the SHAP plot as the reference category shows that participants had 2.64 times the odds (p value < 0.001) of selecting the line diagram and 2.70 times the odds (p value < 0.001) of selecting the bar diagram visualization compared to the SHAP force plot. The SHAP force plot had the worst performance, with an odds ratio of 0.29, indicating a strong preference for other visualizations. Interestingly, participants had 1.20 times the odds (p value $= 0.015$) of selecting options that included the energy-saving tip compared to those that did not.

> Overall, the results of Paper V underscore the importance of aligning state-of-the-art XAI visualizations such as those from the SHAP package with (i) case-specific requirements and (ii) familiar visualizations within the relevant context (e.g., line and bar diagrams for time series data). This became particularly apparent because the SHAP package's force plot performed poorly in both phases of the experiment compared to the other visualizations tested, whereas the line plot performed best. One explanation for this finding might be that time series data is naturally represented on a horizontal time axis, which could be easier for users to understand than a SHAP force plot, highlighting the need for a case-specific tailoring of visualizations. Overall, the findings point to user comprehensibility and preferences as potential factors underlying the effectiveness of XAI-based decision support.

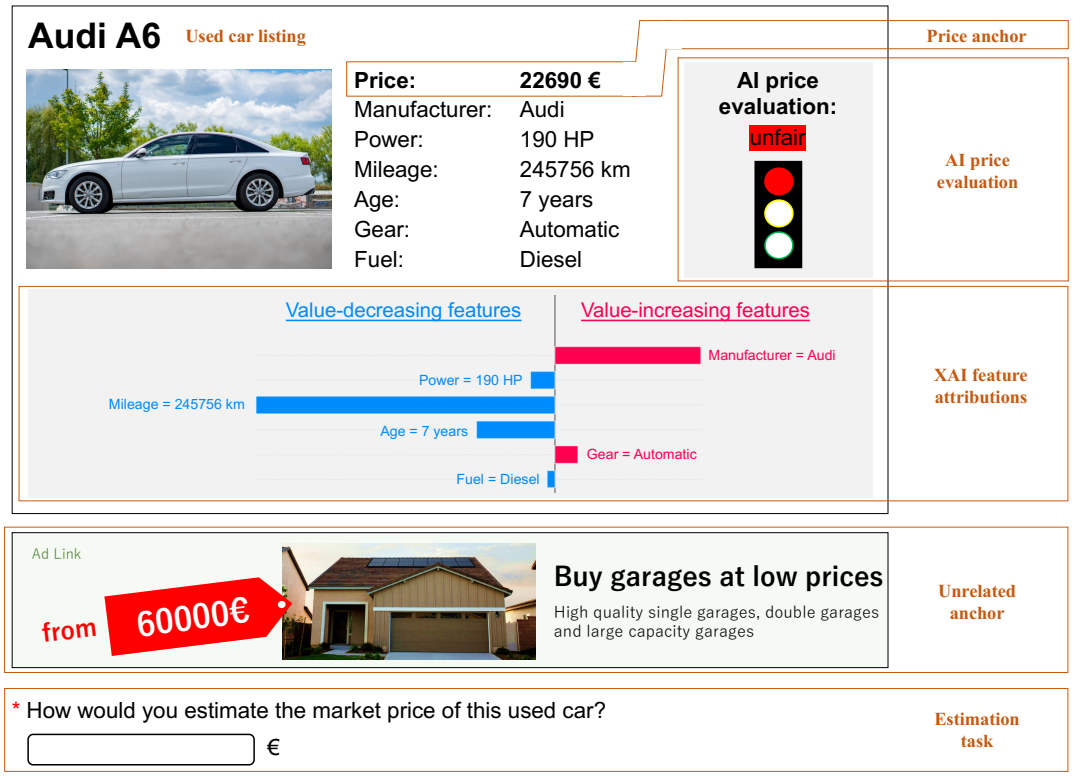
4.2.4 Human-centric evaluation: The utility of XAI-based decision support to reduce anchoring bias (Paper VI*)

As human decision-making often relies on intuition or gut feeling (Tversky and Kahneman, 1974), humans are frequently susceptible to heuristics and cognitive biases (Kahneman, 2011). Of the numerous cognitive biases that have been investigated, anchoring bias is among the most robust of such phenomena (Furnham and Boo, 2011). In IS, platform operators often exploit anchoring bias to prompt behaviors such as impulsive purchases (Moser, 2020). To counteract such negative consequences, DSS literature has tested interventions such as warning messages near anchors (see, e.g., George et al., 2000). The selective accessibility model supports these efforts, suggesting that activating incompatible information with anchors can reduce their impact (Strack et al., 2016). However, the use of AI-based decision support in the form of ML-based predictions and XAI explanations to counter anchoring bias remains largely unexplored and has so far only been conceptually outlined (Wang et al., 2019). In addition, it remains unclear how the salience of XAI explanations affects anchoring bias and, ultimately, human task performance.

Paper VI aims to fill this research gap by testing the effect of AI-based and XAI-based decision support on anchoring bias and task performance. The study took place in the context of the used car market, where participants estimated the market prices of six real used car

∗ F. Haag, C. Stingl, K. Zerfass, K. Hopf, and T. Staake (2023). "Overcoming Anchoring Bias: The Potential of AI and XAI-based Decision Support." *Proceedings of the 44th International Conference on Information Systems*. Hyderabad, India.

listings. The study employed a mixed design, with anchor type (no anchor, price anchor, and unrelated anchor) as a within-subjects factor and decision support as a between-subjects factor. Participants were aided through four support conditions: (i) no support, (ii) AI-based decision support (an implemented ML-based price evaluation that categorized the fairness of an offered price according to the levels unfair, moderate, fair), (iii) XAI-based decision support* (SHAP-based visualization that explained the impact of features on an estimated fair price only), and (iv) combined AI-based and XAI-based decision support (i.e., prediction and explanation). Figure 9 illustrates an example of an estimation task.



Photos: RoClickMag/Shutterstock.com; Vivint Solar/Unsplash.

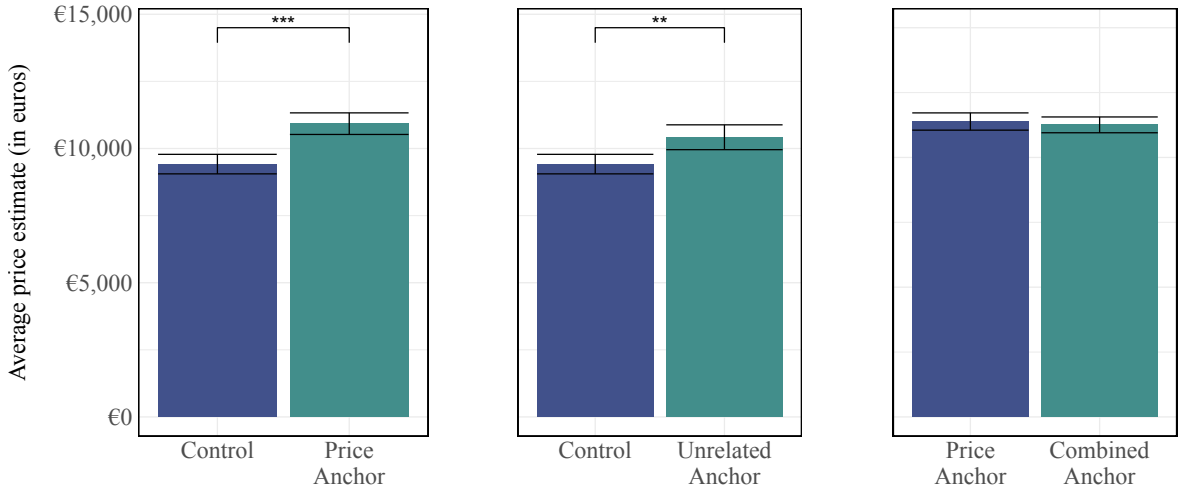
Figure 9: Exemplary illustration of an estimation task (translated from German language; adapted from Haag, Stingl, et al., 2023; Paper VI). This figure is excluded from the CC BY license of this work.

The paper reports the results of two online experiments with N = 390 participants. The first experiment analyzed in a 2×4 factorial design how a price anchor or both an unrelated and a price anchor (within-subject condition) influence price estimations when participants were assisted by one of the four decision support conditions (between-subject condition). The second experiment analyzed the isolated effect of an unrelated anchor as the price anchor from earlier tasks could have affected the subsequent tasks. For this purpose, the second experiment

* In contrast to all other papers, XAI-based decision support in this study refers solely to the explanation, without displaying the corresponding prediction. This approach is necessary to examine the isolated and marginal effect of the explanation as the only aid in decision-making.

used a 2×2 factorial design, with the unrelated anchor as the within-subject condition and no support and XAI-based decision support as the between-subject condition.

The results of this paper are based on three pillars: first, a basic analysis of the occurrence of anchoring bias; second, the effect of the four support conditions on task performance; and third, the effect of these conditions on anchoring bias.



(a) Price anchor (b) Unrelated anchor (c) Combined anchor

Figure 10: Effect of positioned anchors on participants' price estimates. ** and *** indicate statistical significance at the 5% and 1% levels, respectively

With respect to the first pillar, the results highlight the robustness of anchoring bias in the digital space. Specifically, two-sample one-sided Welch's t-tests* determined that both the price anchor (Figure 10a) and the unrelated anchor (Figure 10b) had a statistically significant upward impact on participants' price estimates when compared to estimates without an anchor (+ 15.99% for the price anchor, p value = 0.003; + 10.62% for the unrelated anchor, p value = 0.045). Interestingly, in Experiment 1, when price and unrelated anchors were combined, the resulting price estimates did not differ significantly from those obtained with a price anchor alone (Figure 10c).

The second pillar concerned the effect of the decision support conditions on human task performance. The paper reports task performance by the absolute difference between participants' estimated price and a price considered fair by trained ML models. The corresponding OLS multiple linear regression models** also controlled for the six individual estimation tasks and participants' domain literacy to capture knowledge on cars, car prices, and the automotive

∗ 19 of 390 participants were excluded from the analyses as they made car price estimates that seemed unrealistically low (i.e., below €500).

∗∗ All subsequent reports of p values and effect sizes in percent for Paper VI correspond to estimated regression coefficients rather than direct mean comparisons between groups.

Table 4: Percent reduction in estimation error by decision support condition (relative to the control group)

Support condition	Experiment 1	Experiment 2
Treatment AI	−11.46%	–
Treatment XAI	−12.17%	−15.02%*
Treatment AI & XAI	−16.76%	–

Notes: Percentages are computed based on estimated coefficients from the regression model. * indicates statistical significance of the corresponding coefficient at the 10% level.

market. Table 4 summarizes the relative reduction in estimation error for each decision support condition compared to the baseline condition without decision support.

In Experiment 1, the groups with AI-based and XAI-based decision support, as well as both combined, displayed a lower price estimation error between 11.46% and 16.76% than the control group with no decision aid, although these differences were not statistically significant (Table 4). In Experiment 2, where no price anchors were set, XAI-based decision support improved task performance compared to no decision support at the trend level (-15.02% lower price estimation error, p value = 0.059). More specifically, these results indicate that XAI-based decision support was effective in the absence of a price anchor only. This suggests that XAI-based decision support is effective when no strong external cue (such as a price anchor) competes for attention. In such contexts, the explanations provided by XAI may be more cognitively salient and therefore more likely to guide users' judgments.

Table 5: Deviation from price anchor by support condition and task type (relative to the control group)

Support condition	Fair tasks	Moderate tasks	Unfair tasks
Treatment AI	−35.07%**	−22.92%	+8.62%**
Treatment XAI	−13.06%	−4.82%	−18.98%
Treatment AI & XAI	−48.65%***	−22.26%	+2.13%**

Notes: Percentages are computed based on estimated coefficients from the regression model. ** and *** indicate statistical significance of the corresponding coefficient at the 5% and 1% levels, respectively.

The third pillar of analysis revolves around the effects of the decision support conditions on anchoring bias. The paper measured indications of anchoring as the deviation of participants' price estimates from a displayed price anchor. Table 5 presents the percent change in participants' deviation from the displayed price anchor across different decision support conditions and task types (fair, moderate, unfair) relative to the control group that received no decision support. Specifically, a negative value indicates that participants' estimates were closer to the anchor than those in the control group, while a positive value indicates a greater deviation. Whether a smaller or greater deviation is considered desirable depends on the fairness of the

anchor: in fair tasks, the anchor reflects a reasonable price, meaning that smaller deviations (i.e., negative values relative to the control group) indicate more appropriate estimations. In contrast, in unfair tasks, the anchor is considered misleading, so larger deviations (i.e., positive values relative to the control group) reflect reduced anchoring bias.

The results show that AI-based decision support and its combination with XAI-based decision support helped participants adjust their estimates in the intended direction. In fair tasks, AI-based decision support reduced deviation from the anchor by 35.07% (p value = 0.021), and the combined support by 48.65% (p value = 0.002), compared to the control group, indicating closer alignment with prices that are considered appropriate. For moderate tasks, the analysis also shows a convergence toward the prices displayed for all support conditions, although the results are not significant. In unfair tasks, AI-based decision support increased the deviation by 8.62% (p value = 0.042), and the combined support by 2.13% (p value = 0.023) suggesting that participants were less anchored by misleading prices in comparison to the control group. These results indicate that the support conditions effectively adjusted participants' estimations in accordance with the fairness of the provided price anchor.

> In sum, the results of this paper provide three key insights. First, the salience of explanations appears to be important for effective support, because explanations in the XAI-based decision support conditions only had a significant positive effect on task performance when no price anchor was present (i.e., when no other distorting information was available). Second, with respect to the selective accessibility model, the findings support Mussweiler et al.'s (2000) proposition that anchor-incompatible information can effectively mitigate anchoring effects. Third, the results highlight the importance of a strong fit between the type of decision support and the task at hand. Specifically, only AI-based support and its combination with XAI significantly reduced anchoring bias when appropriate (i.e., when prices were deemed unfair) suggesting that AI-based decision support was the key determinant of effectiveness. One possible explanation for this finding might be the better fit between the task (i.e., estimating prices) and the AI-based decision support (i.e., a traffic-light indicator of price fairness).

4.3 Chapter 3: Instantiating an XAI-based decision support system and evaluating its effects in the field (Papers VII–IX)

The results presented within the scope of Chapter 3 complement the main results of this dissertation by focusing on the instantiation of an XAI-based DSS and the evaluations of its effects on learners in the field. These results address **Research Objective 3**.

4.3.1 Toward personalized feedback that addresses learners' heterogeneity through an XAI-based decision support system (Paper VII*)

Literature has put forth many DSS artifacts that aim to support learning activities and that can have large effects on students' behavior and academic outcomes (see, e.g., Günther, 2021; Theobald and Bellhäuser, 2022). Yet, due to their design, such artifacts often seem to primarily support specific learner groups. For instance, the effect of an intervention that relies on situational achievement goals to support students depends on high prior learning performance and less social activity (Mousavi et al., 2025). One possible explanation for these disparities stems from SRL theory, which posits that learners inherently differ in SRL skills (Credé and Phillips, 2011; Pintrich, 2000). A promising approach to addressing these differences might lie in supervised ML, which has the capability to model the relationships between individual learning trajectories and academic success based on historical data (Hellas et al., 2018). XAI in the form of counterfactuals could then enable the generation of personalized feedback instructions tailored to students' characteristics and platform behavior. Despite the capabilities of supervised ML and XAI, their potential to address individual differences in feedback provision remains largely unexplored.

Paper VII addresses this limitation by instantiating a feedback artifact that leverages ML and XAI to address the heterogeneity of learners in higher education. The research design adopts a nomothetic genre of design science inquiry (Baskerville et al., 2015) by identifying seven key issues for artifact design from established knowledge, namely SRL theory and XAI literature. Building on these issues, the paper describes seven meta-requirements that are then categorized into functional requirements (i.e., what the artifact must do) and non-functional requirements (i.e., measurable properties of the artifact), according to Sommerville (2011).

The subsequent artifact instantiation built on the aforementioned requirements. Figure 11 presents an exemplary feedback output that is embedded into the digital learning environment. To generate such feedback, the artifact initially collected prior data on platform behavior (i.e., platform activities and learning time) and learners' characteristics (e.g., sociodemographic data). The artifact then used these data to train supervised ML models that predict students' academic success (i.e., exam points) for a specific point in time. For a new course run, the artifact passes current data to the ML trained models to predict exam points. Then, *How-to?* explanations and specifically counterfactual explanations (using diverse counterfactual explanations (DiCE) by Mothilal et al. (2020)) can deduce how input features should be altered to shift the model prediction to a desired value. This could, for example, correspond to a change in the model's prediction from 50 to 65 exam points (out of 90), resulting from an increase in the feature values for the number of videos watched and quizzes solved. Finally, the artifact translates feature changes derived from the counterfactual explanations to individual

∗ F. Haag, S. A. Günther, K. Hopf, P. Handschuh, M. Klose, and T. Staake (2023). "Addressing Learners' Heterogeneity in Higher Education: An Explainable AI-based Feedback Artifact for Digital Learning Environments." *Proceedings of the 18th International Conference on Wirtschaftsinformatik*. Paderborn, Germany.

feedback instructions. It is important to note that this approach is solely based on correlations that ML models detect in data. The counterfactual explanations therefore may, but do not necessarily reflect causal real-world relationships for academic success.

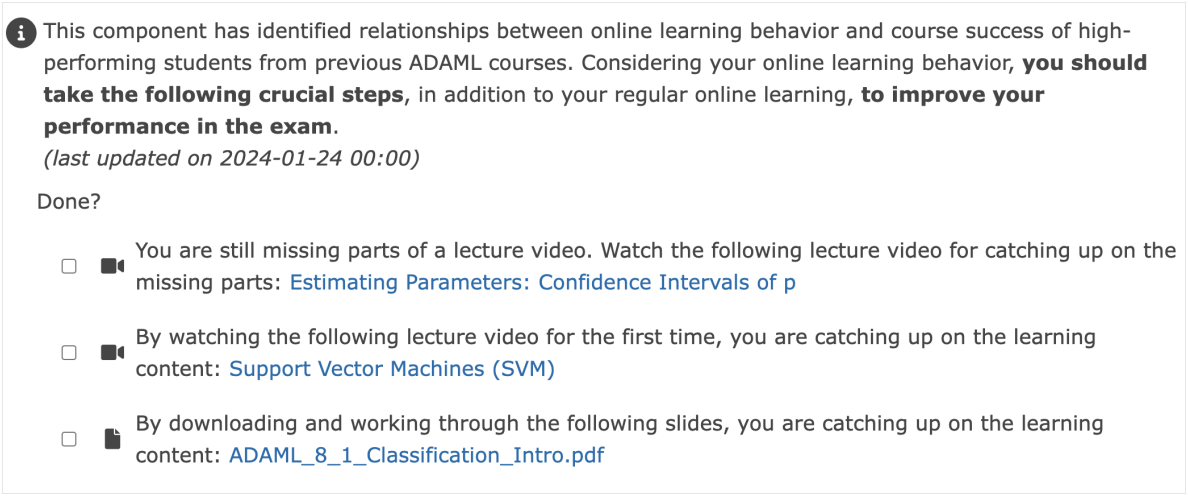


Figure 11: Exemplary feedback for learners generated by the artifact (adapted from Haag, Günther, et al., 2023; Paper VII)

While the functional meta-requirements were inherently considered and fulfilled due to a corresponding development, the paper explicitly reports the test results of the non-functional ones, namely (i) providing feedback that reflects meaningful relationships in the data and (ii) providing feedback that learners find valuable for their learning. The test on whether the feedback reflects meaningful relationships in the data relied on weekly data from three iterations of a data analytics course ($N = 118$) at the University of Bamberg and used predictive performance as a proxy metric. The results of paired one-sided two-sample t-tests on the residuals show that ML estimators significantly outperformed naive estimators (i.e., mean and median) at midterm (week 9: t value = -2.55, p value = 0.009, d value = 0.54) and at the end of the semester (week 17: t value = -3.28, p value = 0.002, d value = 0.67). The assessment on whether learners found the feedback valuable for their individual learning relied on online surveys (N = 36) conducted in two university courses (i.e., a project management course in the summer of 2022 and a data analytics course in the winter of 2022/2023) approximately two weeks before the final exam. The results indicate that learners generally accepted the feedback, perceived it as useful and helpful, and rarely rejected or disputed it. In sum, the paper finds that the artifact (a) effectively models the heterogeneity of learners, and (b) is perceived as valuable by those who received feedback.

In summary, Paper VII instantiates a novel XAI-based DSS that provides feedback and is designed to address differences between individuals. The artifact thereby takes an innovative approach by using XAI not only to explain ML model output for users but as a facility to generate personalized and adaptive feedback. In addition, the paper's approach is also interesting from a theoretical standpoint as it embraces a sociotechnical perspective in XAI feedback design through the inclusion of validated theoretical models from educational psychology. This represents a counter-approach to the sole integration of ML models for predictive analytics and XAI methods to explain model outputs in artifacts that might extend beyond education to other domains. The paper also raises key questions about the effects of such feedback on learning outcomes and whether these effects on learning outcomes are moderated by factors related to learners' SRL (e.g., metacognitive skills or procrastination behavior) or independent of these factors due to the artifact's capabilities to adapt to learners' behavior and characteristics.

4.3.2 A field evaluation of the effects of an XAI-based decision support system for personalized feedback in higher education (Paper VIII* and Paper IX**)

Despite knowing about the benefits of personalized feedback in classroom teaching (Hattie, 2008; Hattie and Timperley, 2007), behavioral interventions in digital learning environments often rely on "one-size-fits-all" support in a content frame that seems to benefit specific learner groups only (see, e.g., Mousavi et al., 2025). In response to this limitation, Paper VII instantiated an XAI-based artifact that aims to provide more personalized feedback in digital learning environments by addressing learners' heterogeneity. Specifically, the approach uses ML to fit models between students' characteristics, platform behavior, and exam performance, and then uses these models to generate counterfactual explanations that suggest individual learning actions based on prior platform behavior.

Paper VIII and Paper IX present a human-centric evaluation of this approach, assessing (i) its effects on students' online learning behavior and (ii) whether factors of SRL moderate these effects. While Paper VIII introduces the research design of the field experiment on a conceptual level, Paper IX reports the results from an actual implementation in an international IT project management course at the University of Bamberg. The field experiment followed a DID experimental design (Figure 12), which began with a six-week baseline phase to collect students' learning behavior data. From course week seven to the exam in course week 19 (i.e., the intervention phase), the treatment group (N = 41) received weekly updated feedback, while

* S. A. Günther, F. Haag, K. Hopf, P. Handschuh, M. Klose, and T. Staake (2023). "A Feedback Component That Leverages Counterfactual Explanations for Smart Learning Support." In: *Digitale Kulturen der Lehre entwickeln – Rahmenbedingungen, Konzepte und Werkzeuge*. Edited by L. Mrohs, J. Franz, D. Herrmann, K. Lindner, and T. Staake. Perspektiven der Hochschuldidaktik. Wiesbaden, Germany: Springer VS.

** F. Haag, S. A. Günther, P. Handschuh, M. Klose, K. Hopf, and T. Staake (submitted). "Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course." Submitted to the European Journal of Information Systems.

the control group (N = 41) received none. After excluding students who had not registered for the exam and did not visit the platform after the baseline phase, N = 32 students remained in the control group and N = 34 in the treatment group. The groups were balanced in key characteristics (e.g., average online study time per week at the end of the baseline phase and self-reported test anxiety).

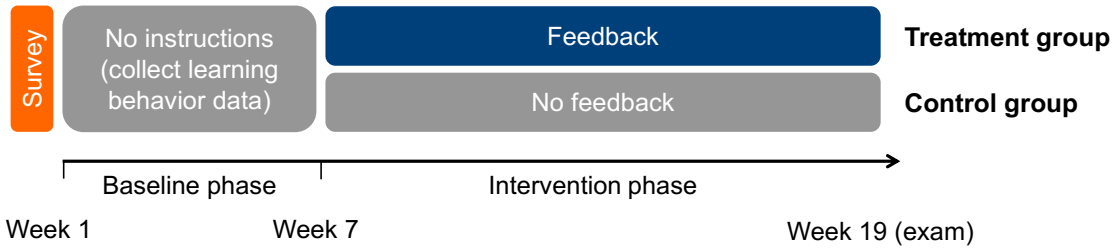


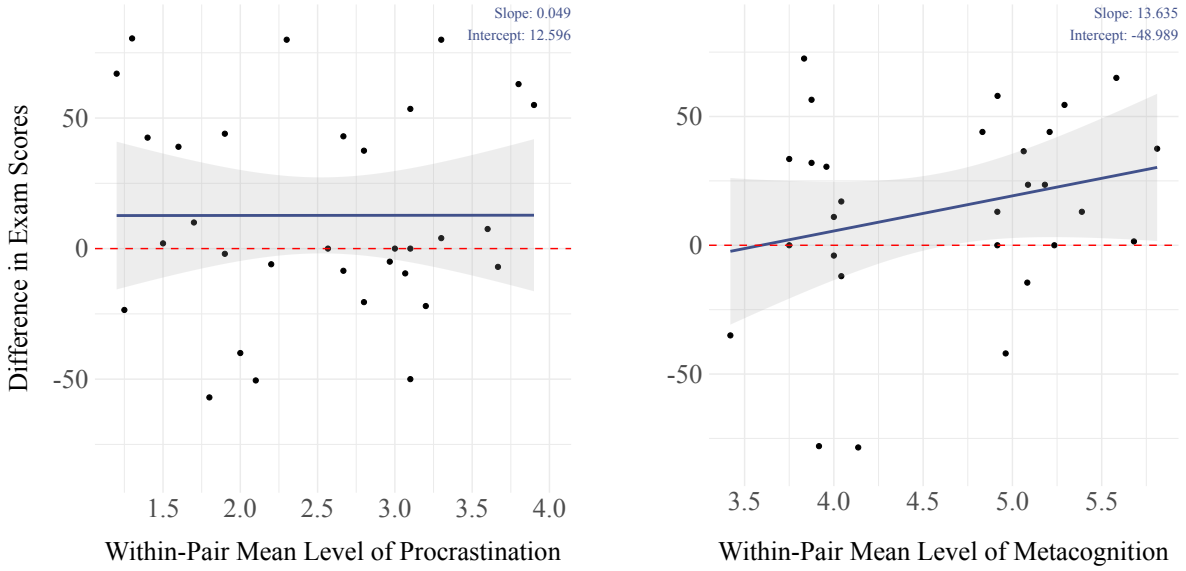
Figure 12: Difference-in-differences design (adapted from Günther et al., 2023; Paper VIII)

The results of Paper IX reveal that treatment group participants increased online learning time, implementation of suggested course content, and exam performance. More specifically, the treatment group spent on average 12.49 minutes more per week on the platform (80.69%) than the control group during the intervention phase (p value = 0.046*). In summary, the treatment effect accumulated 2.71 hours of additional online learning time in the intervention phase. Learners also appeared to follow the feedback instructions (e.g., completing quizzes or watching videos). A two-sided t-test indicated that students in the treatment group implemented significantly more of the respective learning actions than participants in the control group ($p = 0.025$). Students who received feedback seemed to take advantage in terms of the exam score. An ordinary least squares multiple linear regression model showed that the treatment group participants scored 13.18 points higher in the exam (out of 90 points) than participants in the control group (p value = 0.064)**. While participants in the treatment group also displayed a trend toward higher exam participation (2.27 times the odds of participating in the exam compared to the control group), logistic regression analysis found no statistically significant effect of the treatment on the exam-taking rate.

To better understand whether the treatment effects depend on learners' characteristics and behavior, the paper also reports the results of heterogeneity analyses in terms of SRL-related factors and platform usage. To this end, the analysis uses a propensity score matching approach to examine whether the effect of the feedback on exam performance depends on students' self-reported procrastination behavior and the use of metacognitive strategies (i.e., SRL-related variables in the following described as "procrastination" and "metacognition"). Figure 13 shows the relationship between the mean level of the respective SRL-related factors

* Unless stated otherwise, all subsequent reports of p values for Paper IX correspond to statistical significance tests on the estimated regression coefficients rather than direct mean comparisons between groups.

** In line with the approach of the university's examination office, students who have registered for the exam but did not attend received 0 points. Controlling for weeks online yields $\beta = 13.29$, p value = 0.029.



(a) Procrastination

(b) Metacognition

Figure 13: Difference in exam outcomes across SRL factors (adapted from Paper IX)

and the difference in exam score for each pair; the gray area denotes the 95% confidence interval. The plot on procrastination displayed a regression line with slope close to zero and a positive intercept, suggesting that the XAI-based feedback seemed to have benefitted exam performance, independent of procrastination (Figure 13a). Indeed, the moderation analysis found no significant differences between subgroups (implemented as the median split in high and low procrastination behavior), indicating that there is no moderating effect of procrastination (β = -1.01, SE = 15.0, p value = 0.946). Although Figure 13b suggests that higher metacognition might be associated with taking greater advantage of the treatment in terms of exam performance, the corresponding statistical test found no significant difference between groups that were also split in low and high levels by the median (β = 18.8, SE = 14.8, p value = 0.202). Overall, the analyses found no moderating effect of students' metacognition and procrastination levels, which suggests that the XAI-based feedback benefits learners independent of their SRL-related factors.

> The experiment outlined in Paper VIII and implemented in Paper IX yields two main results. First, the implementation of an XAI-based DSS in the form of feedback in an actual university course produced large and statistically significant increases in students' online learning time and in exam performance when controlling for weeks online. Second, the results support that the XAI-based DSS benefits students' academic outcomes, irrespective of their metacognitive skills and procrastination behavior. In summary, these results illustrate the power of personalized feedback and how XAI in the form of counterfactual explanations can enable effective XAI-based decision support.

5 Contributions and practical implications

By examining three research objectives, this dissertation adds to knowledge of the methods, mechanisms, and effects of XAI-based decision support. This chapter outlines the main contributions this dissertation makes to the literature and discusses their practical implications according to the three main chapters.

5.1 Chapter 1: Synthesizing the overall effect of XAI-based decision support on task performance (Paper I and Paper II)

Literature on human-XAI collaboration has reported improvements in decision performance through XAI-based decision support (see, e.g., Bansal et al., 2021; Lai et al., 2020), but has also questioned its effectiveness (see, e.g., Bauer et al., 2023; Schemmer, Hemmer, et al., 2022). Paper I and Paper II (Chapter 1) engage in this scholarly discourse through a comprehensive meta-analysis that synthesizes the overall effect of XAI-based decision support on human task performance. In addition, Paper I and Paper II investigate the factors of the risk of bias in study results and the explanation type provided to better understand their influence on the overall effect estimate. Table 6 presents the main contributions to the literature and translates them into practical implications for the design of DSS.

Table 6: Summary of contributions to the literature and practical implications of Chapter 1

Contributions to the literature	Practical implications
1 Strengthens the evidence that the overall effect of current XAI-based decision support on task performance is small and varies considerably across studies.	A Practitioners might treat the integration of XAI explanations into a DSS as a deliberate choice rather than a default, weighing potential benefits against risks, such as in light of the specific task, domain, and user characteristics.
2 Shows that higher risk of bias in studies is associated with larger reported effects and thereby highlights the importance of accounting for bias in the effect estimation.	B When explanations are adopted, practitioners should consider multiple factors that go beyond the explanation type when designing DSS; this could include aspects such as the explanation quality or salience of explanations.
3 Demonstrates that explanation type alone did not significantly account for variation in task performance, shifting design considerations toward other factors.	

Paper I and Paper II provide three **contributions to the literature**. First, the findings strengthen the evidence that the overall effect of current explanations in XAI-based decision support on task performance is small and varies considerably between studies (1). This provides evidence for the assumption of a stream in the literature that XAI explanations are not universally beneficial for human decision-making (see, e.g., Bauer et al., 2023; Carton et al., 2020). In addition, both papers complement the meta-analysis of Schemmer, Hemmer, et al.

(2022), which included nine articles and reported a small, statistically non-significant difference in task performance between the treatment conditions AI- and XAI-based decision support, possibly due to the limited number of studies available at the time. Hence, by incorporating a total of 27 articles (14 articles in Paper I and the expanded set in Paper II), both papers provide greater confidence in the estimated effect sizes. In sum, this contribution particularly motivates further investigations with respect to the mechanisms that can explain the variation in task performance across studies.

Second, the results show that the risk of bias in study results is positively associated with reported task performance in XAI-based decision support – that is, higher bias tends to coincide with larger reported effects (2). This pattern suggests that the true pooled effect is likely even smaller; when considering only studies with low risk of bias, the estimated effect size approaches zero. Importantly, these results should not be interpreted as evidence that explanations in XAI-based decision support are negligible. Rather, they further underscore the need to investigate the mechanisms driving variation in the effect of such support on task performance. Moreover, by systematically quantifying this bias–effect relationship, this finding demonstrates how studies' methodological quality shapes reported outcomes. This advances the literature by (a) offering a more nuanced interpretation of inconsistent results among studies and (b) pointing to the importance of accounting for bias in meta-analyses.

Third, the findings demonstrate that the explanation type does not account for the variation in task performance among the studies included (3). The literature has highlighted the explanation type as an important factor for human-XAI collaboration (Schauer, 2024), and individual studies indeed report context-dependent differences in the effects of different explanation types (see, e.g., Herm, 2023b). However, when results are interpreted across studies, there is no moderating effect of the explanation type. In other words, there is no evidence of a significant difference between explanation types across the included studies. This finding contributes to the literature by showing that explanation type appears less decisive for decision performance than previously assumed, supporting the notion that a one-size-fits-all approach to the design of XAI-based DSS is unlikely to yield effective support. Overall, it highlights the need to better understand contextual conditions and underlying mechanisms that may constrain when and how explanations in such support exert an effect on human task performance.

Drawing on the findings and contributions outlined above, Chapter 1 formulates two overarching **practical implications** for the design of DSS:

- Careful integration decisions (A): Because the effect of XAI-based decision support on task performance varies and is small when aggregated across current studies, integrating explanations into a DSS should be treated as a deliberate decision rather than as a default. This decision might benefit from weighing the potential benefits against possible downsides (e.g., performance decreases) in light of the specific decision context, domain,

and user characteristics. This shifts the focus from assuming explanations are inherently valuable to ensuring they address a concrete operational need.

- Holistic design (B): When explanations are deemed suitable for a DSS, their design should account for more than just the explanation type. Given that the explanation type alone does not explain variation in task performance, effective designs might require integrating multiple factors, such as explanation quality or salience of explanations among other decision support features. Hence, a design process that combines these factors might better align support with the demands of the specific context.

In sum, Chapter 1 lays an important foundation for research efforts to further understand the mechanisms that account for the variations in the effect of XAI-based decision support on human decision-making.

5.2 Chapter 2: Validating and evaluating explanations in XAI-based decision support (Papers III–VI)

The literature has identified several mechanisms that influence the effectiveness of explanations in XAI-based decision support (see, e.g., Herm, 2023b). Adopting a sociotechnical perspective, Chapter 2 advances this literature stream by investigating mechanisms at two complementary levels: method-centric validation, focusing on the technical foundation of effective human-XAI interaction, and human-centric evaluation, placing the user at the center of the investigation. Together, the papers in this chapter advance our understanding of potential mechanisms underpinning the effectiveness of XAI-based decision support from both a technical and a user-centered standpoint.

The work on **method-centric validation** reported in Paper III and Paper IV jointly **contributes to the literature** in two ways: first by validating explanations through empirical tests of robustness and properties for explanation quality, and second by introducing functionally grounded procedures that allow validation of explanations without human involvement. Table 7 summarizes the main contributions and their practical implications.

Paper III primarily focuses on the validation of explanations with respect to their robustness (1). This validation involves analysis of the temporal persistence of detected patterns – that is, testing whether the same patterns hold when applied to new data from a subsequent year. Using real-world data from a European utility company for the use case of cross-selling prediction, the study demonstrates that SHAP-based explanations are not only valid for one snapshot in time but can also remain robust as the underlying data evolves. Thereby, Paper III also contributes to 2 (Table 7) by introducing a transferable approach to validate explanation robustness without human involvement. The paper complements this approach with an additional analysis that is akin to a property-based stability assessment since it evaluates consistency within subsets

Table 7: Summary of contributions to the literature and practical implications for the work on method-centric validations of Chapter 2

Contributions to the literature	Practical implications
1 Validates explanations on real-world data by testing explanations for pattern robustness and properties for explanation quality (e.g., stability/consistency).	A Encourages practitioners to run pre-deployment tests on explanation patterns for robustness and key properties of explanation quality.
2 Introduces functionally grounded approaches to validate explanations without human involvement.	B Equips practitioners with cost-efficient functionally grounded validations before the integration of explainer methods into DSS.

of observations from the same class. In sum, Paper III advances the literature by demonstrating how XAI explanations can be tested for robustness and by introducing a functionally grounded approach to explanation validation.

Paper IV introduces a functionally grounded approach for explanation validation that works across multiple explainer methods, thereby focusing on feature attribution explanations in numerical prediction tasks (2). This focus contrasts with prior research in this regard, which has primarily focused on classification tasks and typically on a single method type (see, e.g., Alvarez Melis and Jaakkola, 2018; Velmurugan et al., 2021). By demonstrating the approach on real-world data, the study also contributes to 1 (Table 7), showing how explanations can be validated with respect to properties such as stability and consistency. The results reveal that while SHAP generally performs well in this analysis, LIME exhibits weaknesses in certain properties, which is consistent with prior findings from the literature (see, e.g., Velmurugan et al., 2021); in general, post-hoc applied methods such as LIME and SHAP that rely on input perturbations have been criticized in the literature for providing unreliable explanations (Slack et al., 2020). An interpretable ML method (EBM), in contrast, achieves competitive explanation quality while maintaining high predictive accuracy. In sum, Paper IV therefore advances the literature by introducing a broadly applicable validation approach for feature attribution explanations by testing the properties of explanation quality for various methods, as well as by highlighting promising approaches for enhancing understanding of ML model outputs in decision support.

The contributions outlined above with respect to **method-centric validations** translate into two **practical implications** for the implementation of XAI-based DSS:

- Pre-deployment testing of explanation quality (A): The findings show that XAI explanations should not be assumed to be of high quality by default. Before deployment, practitioners are encouraged to systematically test explanation outputs for both pattern robustness and key properties of explanation quality (e.g., stability, consistency). This ensures that explanations remain meaningful as data and contexts evolve and provides

safeguards against misleading explanations. In these tests, one might particularly consider interpretable ML approaches (see, e.g., Kraus et al., 2024; Lou et al., 2013).

- Cost-efficient validation procedures (B): Functionally grounded validation approaches can be applied without human involvement, making it possible to efficiently assess and compare explainer methods in a cost-efficient way prior to integration into a DSS. This allows organizations to filter unsuitable methods early and concentrate human-centric evaluations on explanation techniques that already meet defined standards for explanation quality.

After focusing on contributions to the technical foundation of effective human-XAI interaction, the chapter turns to **human-centric evaluations**. Specifically, the second part of Chapter 2 **contributes to the literature** by enhancing understanding of how explanations in XAI-based decision support affect human decision-making. Whereas Paper V investigates how XAI visualizations shape user comprehensibility and preferences, which could act as mechanisms influencing decision performance, Paper VI concentrates on cognitive biases, more specifically anchoring bias, and the effect of explanation saliency on task performance. In sum, these investigations yield three main contributions to the literature and three corresponding practical implications (Table 8).

Table 8: Summary of contributions to the literature and practical implications for the work on human-centric evaluations of Chapter 2

Contributions to the literature	Practical implications
3 Reveals that XAI visualizations differ considerably in terms of user comprehensibility and preferences (e.g., the SHAP force plot underperformed compared to familiar alternatives).	C DSS interfaces should (re-)align XAI visualizations with familiar formats and data characteristics (e.g., line diagrams for time series data), and tailor them to the decision context rather than relying on package defaults.
4 Demonstrates that AI- and XAI-based decision support can mitigate the negative effects of at least one prevalent cognitive bias (anchoring bias), thereby providing empirical support for a proposition based on the selective accessibility model.	D Practitioners could make use of AI- and XAI-based decision support to reduce undesirable effects of anchoring bias.
5 Suggests explanation saliency as a factor contributing to the effectiveness of XAI-based decision support.	E DSS interfaces should ensure that explanations are sufficiently salient to be effective.

Paper V indicates that user comprehensibility and preferences considerably vary between different XAI visualizations (3). Specifically, the paper employed both objective (e.g., task performance) and subjective measures (e.g., preferred choice of visualizations) to compare the widely used SHAP force plot (see, e.g., Arjunan et al., 2020) with more familiar alternatives such as line and bar diagrams. The results illustrate that the SHAP force plot consistently underperformed across both objective and subjective measures, indicating lower preference and reduced comprehensibility. Interestingly, the line diagram performed comparatively well,

likely due to its natural representation of time series data, a form of presentation particularly well suited to explaining electricity consumption. These findings contribute to the literature on human-XAI interaction by highlighting the importance of carefully choosing and tailoring visualizations in XAI-based decision support.

Paper VI demonstrates that AI- and XAI-based decision support can mitigate the negative effects of a cognitive bias, specifically anchoring bias, thereby providing empirical support for a proposition of the selective accessibility model (4). According to the model, anchoring occurs because exposure to an anchor increases the accessibility of anchor-consistent knowledge, thereby biasing subsequent judgments. Mussweiler et al. (2000) argued that activating anchor-inconsistent knowledge – such as additional disconfirming information – can mitigate anchoring bias. The findings of Paper VI provide empirical support for this proposition. This is particularly noteworthy as Wang et al. (2019) conceptually outlined that AI-based decision support, combined with XAI explanations, could counteract the negative effects of anchoring, but had not empirically tested it. In addition, the paper provides indications that user attention, triggered by explanation saliency, could be a mechanism influencing effectiveness. Specifically, task performance increases through XAI support were only marginally significant when no strong external cue (i.e., the anchor) affected participants' estimates, suggesting that the utility of XAI explanations is shaped by their saliency in the decision environment (5).

Building on the findings and contributions from the **human-centric evaluations** outlined above, Chapter 2 has three overarching **practical implications** for the design and use of XAI-based DSS:

- Contextual tailoring of visualizations (C) – The findings indicate that familiar visualizations such as line and bar charts were comparatively more effective in terms of comprehensibility and user preferences, whereas the SHAP force plot consistently underperformed in both dimensions. For practitioners, this implies that DSS interfaces should not rely on XAI package defaults for explanation visualizations but should instead align visualizations with the data characteristics and decision context.

- Support for mitigating anchoring bias (D) – The results provide empirical evidence that AI- and XAI-based decision support can reduce anchoring effects by activating anchor-incompatible knowledge, thereby lowering biased judgments. Hence, such decision support could help counteract anchoring bias in several practical applications. For instance, the mitigation of anchoring might be particularly relevant for B2C or C2C trading platforms, where fair and appropriate price determination may strengthen the platform's relevance and also benefit buyers and sellers.

- Ensuring XAI explanation saliency (E) – XAI explanations supported task performance only when they were sufficiently salient within the decision environment. When strong external cues such as anchors dominated, XAI explanations had little effect on human task

performance. Consequently, if explanations are intended by the designer to shape user decisions, DSS interfaces should ensure that they are presented clearly and prominently so their benefits can be realized even in complex environments with competing cues.

Overall, Chapter 2 contributes to a deeper understanding of the mechanisms underlying the effectiveness of XAI-based decision support by combining method-centric validations with human-centric evaluations. Taken together, these mechanisms may also provide foundational empirical evidence for inductive theorizing on human-XAI interactions.

5.3 Chapter 3: Instantiating an XAI-based decision support system and evaluating its effects in the field (Papers VII–IX)

While the previous contributions of Chapter 2 on human evaluations stem from controlled study environments, Chapter 3 shifts the focus to investigation of an XAI-based DSS in the field. This investigation addresses an area that has largely been overlooked in XAI-related literature: the application of XAI in naturalistic settings (i.e., application-grounded evaluations) (Brasse et al., 2023). The papers in this chapter advance the literature by presenting such an application of XAI, focusing on two aspects: the instantiation of a novel XAI-based DSS artifact in a real university course (Paper VII), and the evaluation of the effect of such an artifact on students' learning outcomes (Paper VIII and Paper IX). Table 9 summarizes the main contributions of Chapter 3 to the literature and outlines the corresponding implications for practice.

Table 9: Summary of contributions to the literature and practical implications of Chapter 3

Contributions to the literature	Practical implications
1 Demonstrates the feasibility of a novel XAI-based DSS artifact that provides personalized feedback in a university course.	A Policymakers and educators could adopt the XAI-based DSS to deliver more personalized feedback in digital learning environments.
2 Highlights the value of a sociotechnical perspective to DSS instantiation, drawing on SRL theory and linking it with XAI in the form of counterfactual explanations.	B Feedback in XAI-based DSS should be personalized to address the diverse needs of individuals.
3 Provides empirical evidence that such an XAI-based DSS increases learners' learning time and exam performance.	C Other domains such as sales can make use of counterfactual explanations to provide personalized feedback to sales professionals (see the project System One Labs, which is planned to be incorporated as a GmbH and developed into a startup).
4 Shows that learners benefit from XAI-based DSS irrespective of their metacognitive skill or procrastination behavior.	

The papers in this chapter make four main **contributions to the literature** (see Table 9). First, Paper VII demonstrates the feasibility of a novel XAI-based DSS artifact that provides personalized feedback in a university course (1). The artifact shows how ML can capture

individual differences among learners and how XAI, in the form of counterfactual explanations, can translate these insights into actionable instructions. Much like teachers who draw on pedagogical expertise to monitor and respond to learner progress, the artifact identifies patterns in activities and learner characteristics and suggests meaningful next steps. This contribution extends prior work on digital behavioral interventions by moving beyond generic learner support and presents an approach to adaptive, personalized guidance in higher education learning support (see, e.g., Günther et al., 2020).

Second, Paper VII highlights the value of a sociotechnical perspective to DSS instantiation by linking domain-specific theory with XAI in the form of counterfactual explanations (2). Drawing on SRL theory (Pintrich, 2004; Zimmerman and Moylan, 2009), the artifact supports cognitive strategies (e.g., repeating a quiz) across all phases of SRL (e.g., performance phase). Embedding established knowledge obtained from theory into the instantiation process contrasts with a purely technical use of XAI aimed at explaining model predictions. The artifact therefore contributes to IS literature on nomothetic artifact design (see, e.g., Baskerville et al., 2015) and demonstrates how instantiations from a sociotechnical perspective can benefit XAI-based DSS. It also points toward opportunities in other domains – such as workplace productivity or organizational learning – where behavioral theory could similarly inform the instantiation of XAI-based DSS.

Third, the evaluation of the artifact in a real university course provides empirical evidence that an XAI-based DSS can significantly increase learners' online study time and, when controlling for weeks online, their exam performance (3). Learners of the treatment group also implemented significantly more of the instructions than the control group, which shows that the intervention successfully influenced platform behavior as learners have intentionally implemented the respective instructions. These findings show that the anticipated benefits outlined in Paper VII translate into measurable improvements in learning outcomes and thereby extend the discourse on how AI-based behavioral interventions can improve academic success – an area of growing attention in IS research (see, e.g., Tsekouras et al., 2024).

Fourth, Paper IX shows that learners benefit from the XAI-based DSS irrespective of their metacognitive skill or procrastination behavior (4). This result contrasts with much of the existing literature, where behavioral interventions have often been found to disproportionately support specific learners with certain characteristics (see, e.g., Huang et al., 2021; Leung et al., 2023; Mousavi et al., 2025). By demonstrating that counterfactual explanations can adapt feedback to important differences in learners' characteristics, the artifact advances the literature by highlighting how XAI-based DSS can simultaneously improve academic outcomes and reduce educational inequality. More broadly, this underscores the potential of XAI as a foundation for tailoring decision support to diverse user populations in domains where heterogeneity plays a central role.

Expanding upon the findings and contributions outlined above, Chapter 3 has three **practical implications**:

- Adoption in digital education (A): The evaluation demonstrates that the XAI-based DSS increases students' learning time and exam performance irrespective of important learner characteristics. Policymakers and educators could integrate such a DSS into digital learning environments in order to bridge gaps for learners disadvantaged by traditional, one-size-fits-all feedback approaches. The artifact thereby also provides an opportunity for scalable support in digital learning environments in various settings, ranging from education institutions to massive open online courses in the field.
- Personalized feedback to address diverse needs (B): The findings underscore the importance of personalized feedback in addressing the diverse needs of individuals. Practitioners could therefore adopt a sociotechnical perspective for feedback instantiation to support a broad range of individuals in their decisions; this might include drawing on established knowledge from the literature to effectively align system feedback with the characteristics of the respective target group.
- Transfer to other domains (C): Although the artifact was instantiated and evaluated in a university course, the underlying approach of combining ML with counterfactual explanations can be applied in other domains. More specifically, enterprise software providers, such as the project System One Labs (planned to be incorporated as a GmbH and developed into a startup; currently in the pre-foundation stage), could employ this form of XAI-based DSS to deliver personalized feedback tailored to sales professionals.

In sum, Chapter 3 contributes an investigation of an XAI-based DSS in the field, thereby addressing a domain-specific challenge in feedback provision. The chapter bridges the theoretical gap between the heterogeneity of learners postulated by SRL theory and the instantiation of adaptive XAI-based DSS to effectively address these differences. Beyond these contributions, Chapter 3 points to future adoptions of XAI-based DSS in education and in other domains where personalization could enable equitable decision support. In doing so, it positions XAI methods not only as resources to make model predictions more understandable to humans but also as powerful components in DSS to generate value across diverse groups.

6 Limitations and future research

While the research presented in this dissertation offers novel insights, its findings should be interpreted in light of several limitations. This section discusses the three main limitations and outlines opportunities for future research.

The first limitation concerns the reliance on post-hoc XAI methods in this dissertation's experimental research components. While post-hoc methods are advantageous in being applicable across different model classes (Adadi and Berrada, 2018) and interesting to study due to their widespread use (Das and Rad, 2020), they have well-known fundamental issues. Because the explanations of post-hoc methods are only approximations of the underlying ML models, they may fail in certain situations to capture the models' decision-making processes accurately (Kraus et al., 2024; Rudin, 2019). Hence, the use of methods such as SHAP and DiCE could have had an impact on the findings of this dissertation in Chapter 2 and Chapter 3. Moreover, the DiCE-generated explanations applied in Chapter 3 may be especially prone to the Rashomon effect, whereby different yet equally valid counterfactuals exist for the same prediction shift (Molnar, 2025). For instance, DiCE could produce two distinct sets of feature changes that both lead to the same adjustment in exam grade prediction, potentially resulting in divergent feedback for students. This effect might limit the findings as the selected counterfactuals represent only one of several plausible explanations and may not fully capture the range of possible feedback scenarios. Since providing multiple explanations and feedback risks inducing confusion among students, it was necessary to select one set of features to change. To address this issue, this dissertation limited the explanation output and used an algorithmic approach to select three features to alter that translate into actionable feedback instructions (see Paper VII).

Another limitation stems from restriction to specific study settings, populations, and data, which constrains external validity and thus the generalizability of findings. In Chapter 1, this is reflected in the meta-analysis focusing on a reduced set of studies. For instance, the meta-analysis solely includes studies examining task performance in classification tasks, although the literature also comprises studies on numerical prediction tasks (see, for example, Paper VI). Both papers in the chapter chose this focus to reduce between-study heterogeneity in order to prioritize the comparability of the included studies. Nevertheless, the findings of Chapter 1 may be limited in generalizability with respect to the overall effect estimates, as they do not include all studies available in the literature that investigate the effect of XAI-based decision support on human task performance. A further concern of this limitation with respect to Chapter 1 relates to temporal validity: as the body of works grows and the understanding of methods and mechanisms deepens, XAI-based decision support may change, which could substantially alter effect estimates in future meta-analyses. This potential change of the treatment contrasts with meta-analyses in domains such as medicine, where the drug of interest is typically fixed or

subject to slight variations only, resulting in effect estimates that remain comparatively stable over time.

Another concern with respect to the above limitation emerges from the generalizability of the findings beyond the experimental contexts. Chapter 2 has employed controlled experimental environments to ensure higher internal validity and thereby chose domain-specific investigations, yet such designs limit conclusions about how results extend beyond the studied settings. The online experiments may be case-specific, and it is questionable whether the observed mechanisms and effects generalize to other domains and real-world conditions. For instance, the results of Paper VI may not directly transfer to other domains as participants' perceptions of inflated used-car prices and their tendency to revise estimates downward in response to AI- and XAI-based decision support might be specific to the used-car marketplace. In domains such as electrical load forecasting, both upward and downward adjustments of estimations are more meaningful and likely (see, e.g., Giacomazzi et al., 2023), making the observed effects less applicable to other scenarios. Moreover, participants' behavior in real purchasing decisions may differ from that observed in experimental tasks, constraining ecological validity. Similarly, concerns of transferability also arise in method-centric validations, which rely on data from the energy domain and are therefore limited in their applicability to other settings. However, the findings still yield domain-specific insights and uncover important mechanisms for variations in the effectiveness of XAI-based decision support.

A third limitation concerns potential weaknesses with respect to internal validity in the field experiment of Paper IX in Chapter 3. Although field experiments are typically associated with high ecological validity, they often do so at the expense of factors that weaken the ability to attribute observed effects to the treatment. This dissertation has attempted to mitigate these concerns through a DID research design and additionally tested whether there are differences between groups with respect to key characteristics. Nevertheless, it was not practically feasible to capture all sources of bias. For example, confounders such as prior academic performance could not be obtained due to data protection obligations. In addition, interactions between treatment group participants and peers from the control group may have distorted the effects. Specifically, exposure to feedback and specific instructions discussed informally outside the intervention might have spilled over into the control group; this reduces the contrast between treatment and control conditions and may have led to an underestimation of the true effect of the XAI-based DSS on learning outcomes.

Taking into account both the limitations and the contributions of this dissertation, several promising avenues for future research emerge. One promising avenue is the exploration of additional mechanisms underlying the effectiveness of explanations in XAI-based decision support. Chapter 1 analyzed the moderating role of explanation type, but not all planned moderation analyses (e.g., testing whether AI literacy moderates task performance) were feasible due to the limited number of studies reporting the required measures. As more experimental studies become available, such analyses may become possible and provide deeper insights into

the variability of XAI effectiveness across studies in the literature. Similar opportunities arise from the investigations in Chapter 2, which focused on method- and human-centric mechanisms underpinning the effectiveness of explanations in XAI-based decision support. Future work could broaden this line of research by examining additional mechanisms – for instance, through studies on additional cognitive biases using comparable research designs. Taken together, the findings of both chapters offer a solid foundation for future investigations with respect to further mechanisms that might drive the effectiveness of XAI-based decision support.

Another promising avenue for future research stems from Chapter 3 and concerns further investigations of the outlined XAI-based DSS. This dissertation instantiated and evaluated such a system in the context of higher education with a relatively small sample. Future research could extend the scope to other courses with larger sample sizes to address the relatively limited sample here and strengthen the generalizability of the findings. Additionally, it would be valuable to examine whether the effects observed in university courses translate not only to other forms of education, such as organizational training, but also to entirely different domains. For instance, sales offers a promising area, as customer relationship management systems provide the data needed for the instantiation of the outlined XAI-based DSS. Such an instantiation could draw on theories of cognitive biases and behavioral decision-making – for example, by training models on sales outcomes to reveal patterns in customer conversion and generating personalized feedback using counterfactual explanations that could support sales professionals.

A third avenue for future research is the application of interpretable ML. While this dissertation has focused primarily on XAI due to its widespread use (Das and Rad, 2020), interpretable ML represents a promising methodological paradigm that does not require additional approximation methods as the models and their decision-making processes are inherently interpretable. Interestingly, interpretable ML models have already demonstrated predictive performance comparable to black-box ML models (Kraus et al., 2024; Kruschel et al., 2025). However, the inclusion of interpretable ML in experimental studies has so far been limited to only a few investigations (see, e.g., Rosenberger et al., 2025). Interpretable ML and its consideration for experimental studies therefore constitute a promising field for future research from a methodological perspective.

7 Conclusion

This dissertation aimed to advance our understanding of the methods, mechanisms, and effects of XAI-based decision support, with the dual purpose of deepening theoretical insights into human-XAI interaction and supporting the implementation of XAI-based DSS in practice. By conducting research with respect to three overarching research objectives, this dissertation (i) synthesized the overall effect of XAI-based decision support on human task performance, (ii) validated and evaluated explanations in such support, and (iii) evaluated an XAI-based DSS in the field. The findings demonstrated that the overall effect of explanations in XAI-based decision support on human task performance across studies in the literature is small and varies considerably between them. Building on this heterogeneity of effects, this dissertation investigated the potential mechanisms underlying the varying effectiveness of XAI as a source of decision support. The in-depth analyses outlined the temporal persistence of XAI-recognized patterns but also revealed differences in effectiveness for human decision-making with respect to, for example, the type of support provided and the comprehensibility and user preferences for different visualizations. These mechanisms form a theoretical foundation for understanding the effectiveness of human-XAI interaction. In addition, by instantiating and evaluating an XAI-based DSS that provides personalized feedback, this dissertation showed a promising pathway for how XAI can give rise to a new class of DSS. In sum, these findings highlight the potential of XAI-based decision support for human decision-making while also demonstrating its sensitivity to parameterization to attain effectiveness.

From a broader perspective, this dissertation illustrated that the use of explanations for AI outputs is not universally beneficial for decision support; rather, the effective use of XAI explanations requires a nuanced and integrative perspective. XAI may prove to be even more valuable as a system component than as a direct explanatory interface. While this work has advanced understanding of XAI-based decision support, there remain many open questions for future research. Numerous new methods, potential mechanisms, and effects on human decision-making have yet to be investigated, particularly the effects examined in field experiments, which are still scarce. Ultimately, such research could enable XAI to become recognized not only as a tool to foster adoption and fulfill regulatory requirements, but also as a powerful decision support paradigm with desirable implications for individuals, organizations, and society.

References

Adadi, A. and Berrada, M. (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

Adaval, R. and Wyer, R. S. (2011). "Conscious and Nonconscious Comparisons with Price Anchors: Effects on Willingness to Pay for Related and Unrelated Products." *Journal of Marketing Research* 48 (2), pp. 355–365. DOI: 10.1509/jmkr.48.2.355.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, Massachusetts: Harvard Business Review Press. ISBN: 978-1-63369-568-9.

Agrawal, A., Gans, J., and Goldfarb, A. (2023). "Do We Want Less Automation?" *Science* 381 (6654), pp. 155–158. DOI: 10.1126/science.adh9429.

Alvarez Melis, D. and Jaakkola, T. (2018). "Towards Robust Interpretability with Self-Explaining Neural Networks." *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. ISBN: 978-0-691-12035-5.

Ariely, D., Loewenstein, G., and Prelec, D. (2003). ""Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences." *The Quarterly Journal of Economics* 118 (1), pp. 73–106. DOI: 10.1162/00335530360535153.

Arjunan, P., Poolla, K., and Miller, C. (2020). "EnergyStar++: Towards More Accurate and Explanatory Building Energy Benchmarking." *Applied Energy* 276. DOI: 10.1016/j.apenergy.2020.115413.

Arnott, D. and Pervan, G. (2008). "Eight Key Issues for the Decision Support Systems Discipline." *Decision Support Systems* 44 (3), pp. 657–672. DOI: 10.1016/j.dss.2007.09.003.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan. DOI: 10.1145/3411764.3445717.

Baskerville, R. L., Kaul, M., and Storey, V. C. (2015). "Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production." *MIS Quarterly* 39 (3), pp. 541–564. DOI: 10.25300/MISQ/2015/39.3.02.

Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. (2021). "Expl(AI)n It to Me – Explainable AI and Information Systems Research." *Business & Information Systems Engineering* 63 (2), pp. 79–82. DOI: 10.1007/s12599-021-00683-2.

Bauer, K., Von Zahn, M., and Hinz, O. (2023). "Expl(AI)Ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing." *Information Systems Research* 34 (4), pp. 1582–1602. DOI: 10.1287/isre.2023.1199.

Baumeister, R. F. and Vohs, K. D. (2004). *Handbook of Self-Regulation: Research, Theory, and Applications*. New York: Guilford Press. ISBN: 978-1-57230-991-3.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. 1st Edition. Oxford University Press. ISBN: 978-0-19-967811-2.

Bouazizi, S. and Ltifi, H. (2024). "Enhancing Accuracy and Interpretability in EEG-based Medical Decision Making Using an Explainable Ensemble Learning Framework Application for Stroke Prediction." *Decision Support Systems* 178. DOI: 10.1016/j.dss.2023.114126.

Brasse, J., Broder, H. R., Förster, M., Klier, M., and Sigler, I. (2023). "Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions." *Electronic Markets* 33. DOI: 10.1007/s12525-023-00644-5.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3), pp. 199–231. DOI: 10.1214/ss/1009213726.

Bryman, A. (2012). *Social Research Methods*. 4th Edition. Oxford: Oxford Univ. Press. ISBN: 978-0-19-958805-3.

Buchanan, B. and Shortliffe, E. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA, USA: Addison-Wesley. ISBN: 978-0-201-10172-0.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems." *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Cagliari, Italy. DOI: 10.1145/3377325.3377498.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33 (2), pp. 220–239. DOI: 10.1002/bdm.2155.

Canha, D., Kubler, S., Främling, K., and Fagherazzi, G. (2025). "A Functionally-Grounded Benchmark Framework for XAI Methods: Insights and Foundations from a Systematic Literature Review." *ACM Computing Surveys* 57 (12). DOI: 10.1145/3737445.

Carton, S., Mei, Q., and Resnick, P. (2020). "Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity." *Proceedings of the 14th International AAAI Conference on Web and Social Media*. Atlanta, Georgia, USA (virtual). DOI: 10.1609/icwsm.v14i1.7282.

Chen, C., Liaw, A., and Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*. Technical Report 666. University of California, Berkeley.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA, USA: Houghton Mifflin. ISBN: 978-0-395-30790-8.

Coolican, H. (2009). *Research Methods and Statistics in Psychology*. 5th Edition. London: Hodder Education. ISBN: 978-0-340-98344-7.

Cooper, H. M. (2009). *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. 4th Edition. Los Angeles: Sage. ISBN: 978-1-4129-3705-4.

Coussement, K., Abedin, M. Z., Kraus, M., Maldonado, S., and Topuz, K. (2024). "Explainable AI for Enhanced Decision-Making." *Decision Support Systems* 184. DOI: 10.1016/j.dss.2024.114276.

Craven, M. and Shavlik, J. (1995). "Extracting Tree-Structured Representations of Trained Networks." *Proceedings of the 9th International Conference on Neural Information Processing Systems*. Denver, Colorado.

Credé, M. and Phillips, L. A. (2011). "A Meta-Analytic Review of the Motivated Strategies for Learning Questionnaire." *Learning and Individual Differences* 21 (4), pp. 337–346. DOI: 10.1016/j.lindif.2011.03.002.

Croissant, Y. and Millo, G. (2019). *Panel Data Econometrics with R*. Hoboken, NJ, USA: John Wiley & Sons. ISBN: 978-1-119-50464-1.

Das, A. and Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. DOI: 10.48550/arXiv.2006.11371. URL: https://arxiv.org/abs/2006.11371.

Doshi-Velez, F. and Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. DOI: 10.48550/arXiv.1702.08608. URL: http://arxiv.org/abs/1702.08608.

Du, M., Liu, N., Yang, F., Ji, S., and Hu, X. (2019). "On Attribution of Recurrent Neural Network Predictions via Additive Decomposition." *The World Wide Web Conference*. San Francisco, CA, USA. DOI: 10.1145/3308558.3313545.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, pp. 115–118. DOI: 10.1038/nature21056.

Evans, J. S. B. T. (2008). "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59, pp. 255–278. DOI: 10.1146/annurev.psych.59.103006.093629.

Fink, L. (2022). "Why and How Online Experiments Can Benefit Information Systems Research." *Journal of the Association for Information Systems* 23 (6), pp. 1333–1346. DOI: 10.17705/1jais.00787.

Finlay, P. N. (1994). *Introducing Decision Support Systems*. Oxford, UK : Cambridge, Mass., USA: NCC Blackwell ; Blackwell Publishers. ISBN: 978-1-85554-314-0.

Furnham, A. and Boo, H. C. (2011). "A Literature Review of the Anchoring Effect." *The Journal of Socio-Economics* 40 (1), pp. 35–42. DOI: 10.1016/j.socec.2010.10.008.

Fürnkranz, J., Kliegr, T., and Paulheim, H. (2020). "On Cognitive Preferences and the Plausibility of Rule-Based Models." *Machine Learning* 109, pp. 853–898. DOI: 10.1007/s10994-019-05856-5.

George, J. F., Duffy, K., and Ahuja, M. (2000). "Countering the Anchoring and Adjustment Bias With Decision Support Systems." *Decision Support Systems* 29 (2), pp. 195–206. DOI: 10.1016/S0167-9236(00)00074-9.

Giacomazzi, E., Haag, F., and Hopf, K. (2023). "Short-Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources." *Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy)*. Orlando, FL, USA. DOI: 10.1145/3575813.3597345.

Gregor, S. and Benbasat, I. (1999). "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice." *MIS Quarterly* 23 (4), pp. 497–530. DOI: 10.2307/249487.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5). DOI: 10.1145/3236009.

Gunning, D., Vorm, E., Wang, J. Y., and Turek, M. (2021). "DARPA's Explainable AI (XAI) Program: A Retrospective." *Applied AI Letters* 2 (4). DOI: 10.1002/ail2.61.

Günther, S. A. (2021). "The Impact of Social Norms on Students' Online Learning Behavior: Insights from Two Randomized Controlled Trials." *Proceedings of the 11th International Conference on Learning Analytics & Knowledge*. Irvine, CA, USA. DOI: 10.1145/3448139.3448141.

Günther, S. A., Haag, F., Hopf, K., Handschuh, P., Klose, M., and Staake, T. (2023). "A Feedback Component That Leverages Counterfactual Explanations for Smart Learning Support." In: *Digitale Kulturen der Lehre entwickeln – Rahmenbedingungen, Konzepte und Werkzeuge*. Edited by L. Mrohs, J. Franz, D. Herrmann, K. Lindner, and T. Staake. Perspektiven der Hochschuldidaktik. Wiesbaden, Germany: Springer VS.

Günther, S. A., Veihelmann, T., and Staake, T. (2020). "Leveraging Social Norms to Encourage Online Learning: Empirical Evidence from a Blended Learning Course." *Proceedings of the 41st International Conference on Information Systems*. Hyderabad, India (virtual).

Haag, F. (2025). "The Effect of Explainable AI-based Decision Support on Human Task Performance: A Meta-Analysis." *Proceedings of the 23rd Annual Pre-ICIS Workshop on HCI Research in MIS*. Bangkok, Thailand.

Haag, F. (2026). "How Explanations from XAI-based Decision Support Affect Human Task Performance: A Meta-Analysis." *Journal of Decision Systems* 35 (1). DOI: 10.1080/12460125.2026.2616693.

Haag, F., Günther, S. A., Handschuh, P., Klose, M., Hopf, K., and Staake, T. (submitted). "Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course." Submitted to the European Journal of Information Systems.

Haag, F., Günther, S. A., Hopf, K., Handschuh, P., Klose, M., and Staake, T. (2023). "Addressing Learners' Heterogeneity in Higher Education: An Explainable AI-based Feedback Artifact for Digital Learning Environments." *Proceedings of the 18th International Conference on Wirtschaftsinformatik*. Paderborn, Germany.

Haag, F., Hopf, K., Menelau Vasconcelos, P., and Staake, T. (2022). "Augmented Cross-Selling Through Explainable AI—A Case From Energy Retailing." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Haag, F., Hopf, K., and Staake, T. (2026). "Validating Explainer Methods: A Functionally Grounded Approach for Numerical Forecasting." *Journal of Forecasting* 45 (2), pp. 819–836. DOI: 10.1002/for.70060.

Haag, F., Stingl, C., Zerfass, K., Hopf, K., and Staake, T. (2023). "Overcoming Anchoring Bias: The Potential of AI and XAI-based Decision Support." *Proceedings of the 44th International Conference on Information Systems*. Hyderabad, India.

Harrer, M., Cuijpers, P., Furukawa, T., and Ebert, D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. 1st Edition. Boca Raton, FL and London: Chapman & Hall/CRC Press. ISBN: 978-0-367-61007-4.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd Edition. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84858-7.

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. London: Routledge. ISBN: 978-1-134-02412-4.

Hattie, J. and Timperley, H. (2007). "The Power of Feedback." *Review of Educational Research* 77 (1), pp. 81–112. DOI: 10.3102/003465430298487.

Hedges, L. V. (1981). "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics* 6 (2), pp. 107–128. DOI: 10.2307/1164588.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S. N. (2018). "Predicting Academic Performance: A Systematic Literature Review." *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. Larnaca, Cyprus. DOI: 10.1145/3293881.3295783.

Hemmer, P., Schemmer, M., Riefle, L., Rosellen, N., Vössing, M., and Kühl, N. (2022). "Factors That Influence the Adoption of Human-AI Collaboration in Clinical Decision-Making." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Herm, L.-V. (2023a). "Algorithmic Decision-Making Facilities: Perception and Design of Explainable AI-based Decision Support Systems." PhD thesis. Universität Würzburg.

Herm, L.-V. (2023b). "Impact of Explainable AI on Cognitive Load: Insights from an Empirical Study." *Proceedings of the 31st European Conference on Information Systems*. Kristiansand, Norway.

Herm, L.-V., Wanner, J., Seubert, F., and Janiesch, C. (2021). "I Don't Get It, but It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research." *Proceedings of the 29th European Conference on Information Systems*. Marrakech, Morocco (virtual).

Hevner, A., March, S. T., Park, J., and Ram, S. (2004). "Design Science in Information Systems Research." *MIS Quarterly* 28 (1), pp. 75–105. DOI: 10.2307/25148625.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A., eds. (2024). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5 (updated August 2024). URL: https://www.cochrane.org/handbook.

Huang, N., Zhang, J., Burtch, G., Li, X., and Chen, P. (2021). "Combating Procrastination on Massive Online Open Courses via Optimal Calls to Action." *Information Systems Research* 32 (2), pp. 301–317. DOI: 10.1287/isre.2020.0974.

Imbens, G. W. and Wooldridge, J. M. (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1), pp. 5–86. DOI: 10.1257/jel.47.1.5.

Jacowitz, K. E. and Kahneman, D. (1995). "Measures of Anchoring in Estimation Tasks." *Personality and Social Psychology Bulletin* 21 (11), pp. 1161–1166. DOI: 10.1177/01461672952111004.

Johnson, F., Pratt, M., and Wardle, J. (2012). "Dietary Restraint and Self-Regulation in Eating Behavior." *International Journal of Obesity* 36, pp. 665–674. DOI: 10.1038/ijo.2011.156.

Kahneman, D. (2011). *Thinking, Fast and Slow*. 1st Edition. New York: Farrar, Straus and Giroux. ISBN: 978-0-374-27563-1.

King, W. R. and He, J. (2005). "Understanding the Role and Methods of Meta-Analysis in IS Research." *Communications of the Association for Information Systems* 16, pp. 665–686. DOI: 10.17705/1CAIS.01632.

Kizilcec, R. F., Pérez-Sanagustín, M., and Maldonado, J. J. (2017). "Self-Regulated Learning Strategies Predict Learner Behavior and Goal Attainment in Massive Open Online Courses." *Computers & Education* 104, pp. 18–33. DOI: 10.1016/j.compedu.2016.10.001.

Klayman, J. and Ha, Y.-w. (1987). "Confirmation, Disconfirmation, and Information in Hypothesis Testing." *Psychological Review* 94 (2), pp. 211–228. DOI: 10.1037/0033-295X.94.2.211.

Kraus, M., Tschernutter, D., Weinzierl, S., and Zschech, P. (2024). "Interpretable Generalized Additive Neural Networks." *European Journal of Operational Research* 317 (2), pp. 303–316. DOI: 10.1016/j.ejor.2023.06.032.

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., and Zschech, P. (2025). "Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models." *Business & Information Systems Engineering*. DOI: 10.1007/s12599-024-00922-2.

Kühl, N., Hirt, R., Baier, L., Schmitz, B., and Satzger, G. (2021). "How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine

Learning Report Card." *Communications of the Association for Information Systems* 48, pp. 589–615. DOI: 10.17705/1CAIS.04845.

Lai, V., Liu, H., and Tan, C. (2020). "'Why Is "Chicago" Deceptive?' Towards Building Model-Driven Tutorials for Humans." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA. DOI: 10.1145/3313831.3376873.

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W., and Yue, W. T. (2023). "Could Gamification Designs Enhance Online Learning Through Personalization? Lessons from a Field Experiment." *Information Systems Research* 34 (1), pp. 27–49. DOI: 10.1287/isre.2022.1123.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). "Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems." *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*. Boston, MA, USA. DOI: 10.1145/1518701.1519023.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). "Accurate Intelligible Models with Pairwise Interactions." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago, IL, USA. DOI: 10.1145/2487575.2487579.

Lundberg, S. M. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions." *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, USA.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Madsen, A., Reddy, S., and Chandar, S. (2023). "Post-Hoc Interpretability for Neural NLP: A Survey." *ACM Computing Surveys* 55 (8). DOI: 10.1145/3546577.

Maier, U. and Klotz, C. (2022). "Personalized Feedback in Digital Learning Environments: Classification Framework and Literature Review." *Computers and Education: Artificial Intelligence* 3. DOI: 10.1016/j.caeai.2022.100080.

Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., Gabel, S., Goethals, S., Greene, T., Klein, N., Kraus, M., Kühl, N., Perlich, C., Verbeke, W., Zharova, A., Zschech, P., and Provost, F. (2025). *Beware of "Explanations" of AI*. DOI: 10.48550/arXiv.2504.06791. URL: http://arxiv.org/abs/2504.06791.

Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2020). "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities." *Information Systems Management* 39 (1), pp. 53–63. DOI: 10.1080/10580530.2020.1849465.

Miller, T. (2019). "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267, pp. 1–38. DOI: 10.1016/j.artint.2018.07.007.

Miró-Nicolau, M., Jaume-i-Capó, A., and Moyà-Alcover, G. (2025). "A Comprehensive Study on Fidelity Metrics for XAI." *Information Processing & Management* 62 (1). DOI: 10.1016/j.ipm.2024.103900.

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." *ACM Transactions on Interactive Intelligent Systems* 11 (3–4). DOI: 10.1145/3387166.

Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3rd Edition. ISBN: 978-3-911578-03-5. URL: https://christophm.github.io/interpretable-ml-book.

Moser, C. (2020). "Impulse Buying: Designing for Self-Control with E-Commerce." PhD thesis. University of Michigan.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain. DOI: 10.1145/3351095.3372850.

Mousavi, N., Golara, S., and Bockstedt, J. (2025). "The Impact of Situational Achievement Goals on Online Learning Behavior: Results from Field Experiments." *Information Systems Research* 36 (2), pp. 983–1010. DOI: 10.1287/isre.2022.0353.

Mussweiler, T. and Strack, F. (1999). "Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring." *European Review of Social Psychology* 10 (1), pp. 135–167. DOI: 10.1080/14792779943000044.

Mussweiler, T., Strack, F., and Pfeiffer, T. (2000). "Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility." *Personality and Social Psychology Bulletin* 26 (9), pp. 1142–1150. DOI: 10.1177/01461672002611010.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. (2023). "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." *ACM Computing Surveys* 55 (13s). DOI: 10.1145/3583558.

Northcraft, G. B. and Neale, M. A. (1987). "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions." *Organizational Behavior and Human Decision Processes* 39 (1), pp. 84–97. DOI: 10.1016/0749-5978(87)90046-X.

O'Keefe, R. M. and O'Leary, D. E. (1993). "Expert System Verification and Validation: A Survey and Tutorial." *Artificial Intelligence Review* 7, pp. 3–42. DOI: 10.1007/BF00849196.

Panadero, E. (2017). "A Review of Self-regulated Learning: Six Models and Four Directions for Research." *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00422.

Paré, G. and Kitsiou, S. (2017). "Chapter 9 Methods for Literature Reviews." In: *Handbook of eHealth Evaluation: An Evidence-Based Approach [Internet]*. University of Victoria. URL: https://www.ncbi.nlm.nih.gov/books/NBK481583/.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24 (3), pp. 45–77. DOI: 10.2753/MIS0742-1222240302.

Pintrich, P. R. (2000). "The Role of Goal Orientation in Self-Regulated Learning." In: *Handbook of Self-Regulation*. Elsevier, pp. 451–502. DOI: 10.1016/B978-012109890-2/50043-3.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor, MI, USA: National Center for Research to Improve Postsecondary Teaching and Learning. URL: https://eric.ed.gov/?id=ED338122.

Pintrich, P. R. (2004). "A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students." *Educational Psychology Review* 16 (4), pp. 385–407. DOI: 10.1007/s10648-004-0006-x.

Recker, J. (2021). *Scientific Research in Information Systems: A Beginner's Guide*. Progress in IS. Cham: Springer International Publishing. ISBN: 978-3-030-85436-2.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA. DOI: 10.1145/2939672.2939778.

Richardson, M., Abraham, C., and Bond, R. (2012). "Psychological Correlates of University Students' Academic Performance: A Systematic Review and Meta-Analysis." *Psychological Bulletin* 138 (2), pp. 353–387. DOI: 10.1037/a0026838.

Robnik-Šikonja, M. and Bohanec, M. (2018). "Perturbation-Based Explanations of Prediction Models." In: *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Cham: Springer International Publishing, pp. 159–175. DOI: 10.1007/978-3-319-90403-0_9.

Rosenberger, J., Schröppel, P., Kruschel, S., Kraus, M., Zschech, P., and Förster, M. (2025). "Navigating the Rashomon Effect: How Personalization Can Help Adjust Interpretable Machine Learning Models to Individual Users." *Proceedings of the 33rd European Conference on Information Systems*. Amman, Jordan.

Rudin, C. (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Sarker, S., Chatterjee, S., Xiao, X., and Elbanna, A. (2019). "The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and Its Continued Relevance." *MIS Quarterly* 43 (3), pp. 695–719. DOI: 10.25300/misq/2019/13747.

Schauer, A. (2024). "In the Eyes of the User: A Systematic Literature Review of User Perceptions in Human-XAI Interaction." *Proceedings of the 45th International Conference on Information Systems*. Bangkok, Thailand.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. (2022). "A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. Oxford, United Kingdom. DOI: 10.1145/3514094.3534128.

Schemmer, M., Kühl, N., Benz, C., and Satzger, G. (2022). "On the Influence of Explainable AI on Automation Bias." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., and Keim, D. A. (2019). "Towards A Rigorous Evaluation Of XAI Methods On Time Series." *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South). DOI: 10.1109/ICCVW.2019.00516.

Schraw, G., Wadkins, T., and Olafson, L. (2007). "Doing the Things We Do: A Grounded Theory of Academic Procrastination." *Journal of Educational Psychology* 99 (1), pp. 12–25. DOI: 10.1037/0022-0663.99.1.12.

Schwalbe, G. and Finzel, B. (2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts." *Data Mining and Knowledge Discovery* 38 (5), pp. 3043–3101. DOI: 10.1007/s10618-022-00867-8.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2010). "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40 (1), pp. 185–197. DOI: 10.1109/TSMCA.2009.2029559.

Senécal, C., Koestner, R., and Vallerand, R. J. (1995). "Self-Regulation and Academic Procrastination." *The Journal of Social Psychology* 135 (5), pp. 607–619. DOI: 10.1080/00224545.1995.9712234.

Senoner, J., Netland, T., and Feuerriegel, S. (2022). "Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing." *Management Science* 68 (8), pp. 5704–5723. DOI: 10.1287/mnsc.2021.4190.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, USA: Houghton Mifflin.

Shin, D. (2021). "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI." *International Journal of Human-Computer Studies* 146. DOI: 10.1016/j.ijhcs.2020.102551.

Shmueli, G. (2010). "To Explain or to Predict?" *Statistical Science* 25 (3), pp. 289–310. DOI: 10.1214/10-STS330.

Shmueli, G. and Koppius, O. R. (2011). "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35 (3), pp. 553–572. DOI: 10.2307/23042796.

Shmueli, G., Martens, D., Yoo, J., and Greene, T. (2025). *From What Ifs to Insights: Counterfactuals in Causal Inference vs. Explainable AI*. DOI: 10.48550/arXiv.2505.13324. URL: https://arxiv.org/abs/2505.13324.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and

Hassabis, D. (2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529, pp. 484–489. DOI: 10.1038/nature16961.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA. DOI: 10.1145/3375627.3375830.

Sommerville, I. (2011). *Software Engineering*. 9th Edition. Boston: Pearson. ISBN: 978-0-13-703515-1.

Song, W. and Qian, X. (2020). "Adverse Childhood Experiences and Teen Sexual Behaviors: The Role of Self-Regulation and School-Related Factors." *Journal of School Health* 90 (11), pp. 830–841. DOI: 10.1111/josh.12947.

Steel, P. (2007). "The Nature of Procrastination: A Meta-Analytic and Theoretical Review of Quintessential Self-Regulatory Failure." *Psychological Bulletin* 133 (1), pp. 65–94. DOI: 10.1037/0033-2909.133.1.65.

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Reeves, B. C., Shepperd, S., Shrier, I., Stewart, L. A., Tilling, K., White, I. R., Whiting, P. F., and Higgins, J. P. T. (2019). "RoB 2: A Revised Tool for Assessing Risk of Bias in Randomised Trials." *BMJ*. DOI: 10.1136/bmj.l4898.

Stingl, C., Günther, S. A., and Staake, T. (2022). "The Behavioral Mechanisms Behind Feedback – A Preliminary Model for Quantifying Cause-Effect Relationships." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Strack, F., Bahník, Š., and Mussweiler, T. (2016). "Anchoring: Accessibility as a Cause of Judgmental Assimilation." *Current Opinion in Psychology* 12, pp. 67–70. DOI: 10.1016/j.copsyc.2016.06.005.

Štrumbelj, E., Kononenko, I., and Robnik-Šikonja, M. (2009). "Explaining Instance Classifications with Interactions of Subsets of Feature Values." *Data & Knowledge Engineering* 68 (10), pp. 886–904. DOI: 10.1016/j.datak.2009.01.004.

Sturm, T., Pumplun, L., Gerlach, J. P., Kowalczyk, M., and Buxmann, P. (2023). "Machine Learning Advice in Managerial Decision-Making: The Overlooked Role of Decision Makers' Advice Utilization." *The Journal of Strategic Information Systems* 32 (4). DOI: 10.1016/j.jsis.2023.101790.

Suh, A., Hurley, I., Smith, N., and Siu, H. C. (2025). "Fewer Than 1% of Explainable AI Papers Validate Explainability with Humans." *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan. DOI: 10.1145/3706599.3719964.

Swartout, W. R. (1983). "XPLAIN: A System for Creating and Explaining Expert Consulting Programs." *Artificial Intelligence* 21 (3), pp. 285–325. DOI: 10.1016/S0004-3702(83)80014-9.

Templier, M. and Paré, G. (2018). "Transparency in Literature Reviews: An Assessment of Reporting Practices Across Review Types and Genres in Top IS Journals." *European Journal of Information Systems* 27 (5), pp. 503–550. DOI: 10.1080/0960085X.2017.1398880.

Theobald, M. and Bellhäuser, H. (2022). "How Am I Going and Where to Next? Elaborated Online Feedback Improves University Students' Self-Regulated Learning and Performance." *The Internet and Higher Education* 55. DOI: 10.1016/j.iheduc.2022.100872.

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." *Management Science* 64 (3), pp. 1458–1476. DOI: 10.1287/mnsc.2016.2646.

Tsekouras, D., Belo, R., and Cornelius, P. (2024). "Generative AI and Student Performance: Evidence from a Large-Scale Intervention." *Proceedings of the 45th International Conference on Information Systems*. Bangkok, Thailand.

Tuckman, B. W. (2005). "Relations of Academic Procrastination, Rationalizations, and Performance in a Web Course With Deadlines." *Psychological Reports* 96 (3), pp. 1015–1021. DOI: 10.2466/pr0.96.3c.1015-1021.

Tversky, A. and Kahneman, D. (1974). "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (4157), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

Van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). "Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations." *Artificial Intelligence* 291. DOI: 10.1016/j.artint.2020.103404.

Van Lent, M., Fisher, W., and Mancuso, M. (2004). "An Explainable Artificial Intelligence System for Small-unit Tactical Behavior." *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence*. San Jose, California.

Velmurugan, M., Ouyang, C., Moreira, C., and Sindhgatta, R. (2021). "Evaluating Stability of Post-hoc Explanations for Business Process Predictions." In: *Service-Oriented Computing*. Volume 13121. Cham: Springer International Publishing, pp. 49–64. DOI: 10.1007/978-3-030-91431-8_4.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques." *Expert Systems with Applications* 38 (3), pp. 2354–2364. DOI: 10.1016/j.eswa.2010.08.023.

Vosniadou, S. (2020). "Bridging Secondary and Higher Education. The Importance of Self-Regulated Learning." *European Review* 28 (S1), pp. 94–103. DOI: 10.1017/S1062798720000939.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). "Designing Theory-Driven User-Centric Explainable AI." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland. DOI: 10.1145/3290605.3300831.

Wang, P. and Ding, H. (2024). "The Rationality of Explanation or Human Capacity? Understanding the Impact of Explainable Artificial Intelligence on Human-AI Trust and Decision Performance." *Information Processing & Management* 61 (4). DOI: 10.1016/j.ipm.2024.103732.

Wang, W. and Benbasat, I. (2007). "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs." *Journal of Management Information Systems* 23 (4), pp. 217–246. DOI: 10.2753/MIS0742-1222230410.

Wang, X. and Yin, M. (2021). "Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making." *Proceedings of the International Conference on Intelligent User Interfaces*. College Station, TX, USA. DOI: 10.1145/3397481.3450650.

Wanner, J., Heinrich, K., Janiesch, C., and Zschech, P. (2020). "How Much AI Do You Require? Decision Factors for Adopting AI Technology." *Proceedings of the 41st International Conference on Information Systems*. Hyderabad, India (virtual).

Wäschle, K., Allgaier, A., Lachner, A., Fink, S., and Nückles, M. (2014). "Procrastination and Self-Efficacy: Tracing Vicious and Virtuous Circles in Self-Regulated Learning." *Learning and Instruction* 29, pp. 103–114. DOI: 10.1016/j.learninstruc.2013.09.005.

Wastensteiner, J., Weiss, T. M., Haag, F., and Hopf, K. (2021). "Explainable AI for Tailored Electricity Consumption Feedback – An Experimental Evaluation of Visualizations." *Proceedings of the 29th European Conference on Information Systems*. Marrakech, Morocco (virtual).

Weber, R. O., Johs, A. J., Goel, P., and Silva, J. M. (2024). "XAI Is in Trouble." *AI Magazine* 45 (3), pp. 300–316. DOI: 10.1002/aaai.12184.

Wild, K.-P. and Schiefele, U. (1994). "Lernstrategien Im Studium: Ergebnisse Zur Faktorenstruktur Und Reliabilität Eines Neuen Fragebogens." *Zeitschrift für Differentielle und Diagnostische Psychologie* 15 (4), pp. 185–200.

Wilke, A. and Mata, R. (2012). "Cognitive Bias." In: *Encyclopedia of Human Behavior (Second Edition)*. San Diego: Academic Press, pp. 531–535. DOI: 10.1016/B978-0-12-375000-6.00094-X.

Wilson, H. J. and Daugherty, P. R. (2018). "Collaborative Intelligence: Humans and AI Are Joining Forces." *Harvard Business Review* 96, pp. 114–123.

Wolters, C. A. (2003). "Understanding Procrastination from a Self-Regulated Learning Perspective." *Journal of Educational Psychology* 95 (1), pp. 179–187. DOI: 10.1037/0022-0663.95.1.179.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. 6th Edition. Boston: Cengage Learning. ISBN: 978-1-305-27010-7.

Wu, C.-S. and Cheng, F.-F. (2011). "The Joint Effect of Framing and Anchoring on Internet Buyers' Decision-Making." *Electronic Commerce Research and Applications* 10 (3), pp. 358–368. DOI: 10.1016/j.elerap.2011.01.002.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). "Making Sense of Recommendations." *Journal of Behavioral Decision Making* 32 (4), pp. 403–414. DOI: 10.1002/bdm.2118.

Zimmerman, B. J. and Moylan, A. R. (2009). "Self-Regulation: Where Metacognition and Motivation Intersect." In: *Handbook of Metacognition in Education*. New York: Routledge, pp. 299–315.

Chapter 1: Synthesizing the overall effect of XAI-based decision support on task performance

Paper I

The Effect of Explainable AI-based Decision Support on Human Task Performance: A Meta-Analysis

Felix Haag
University of Bamberg

Proceedings of the 23rd Annual Pre-ICIS Workshop on HCI Research in MIS, Bangkok, Thailand, 2025, https://aisel.aisnet.org/sighci2024/17/

Paper II

How Explanations from XAI-based Decision Support Affect Human Task Performance: A Meta-Analysis

Felix Haag
University of Bamberg

Journal of Decision Systems 35 (1), Article nr. 2616693, 2026, https://doi.org/10.1080/12460125.2026.2616693

Chapter 2: Validating and evaluating explanations in XAI-based decision support

Paper III

Augmented Cross-Selling Through Explainable AI—A Case From Energy Retailing

Felix Haag
University of Bamberg

Konstantin Hopf
University of Bamberg

Pedro Menelau Vasconcelos
University of Bamberg

Thorsten Staake
University of Bamberg

Proceedings of the 30th European Conference on Information Systems, Paper nr. 1722, Timișoara, Romania, 2022, https://aisel.aisnet.org/ecis2022_rp/129/

Paper IV

Validating Explainer Methods: A Functionally Grounded Approach for Numerical Forecasting

Felix Haag
University of Bamberg

Konstantin Hopf
University of Bamberg

Thorsten Staake
University of Bamberg

Journal of Forecasting 45 (2), pp. 819-836, 2026, https://doi.org/10.1002/for.70060

Paper V

Explainable AI for Tailored Electricity Consumption Feedback – An Experimental Evaluation of Visualizations

Jacqueline Wastensteiner
University of Bamberg

Tobias Michael Weiss
University of Bamberg

Felix Haag
University of Bamberg

Konstantin Hopf
University of Bamberg

Proceedings of the 29th European Conference on Information Systems, Paper nr. 1284, Marrakech, Morocco (virtual), 2021, https://aisel.aisnet.org/ecis2021_rp/55/

Paper VI

Overcoming Anchoring Bias: The Potential of AI and XAI-based Decision Support

Felix Haag
University of Bamberg

Carlo Stingl
University of Bamberg

Katrin Zerfass
University of Bamberg

Konstantin Hopf
University of Bamberg

Thorsten Staake
University of Bamberg

Proceedings of the 44th International Conference on Information Systems, Paper nr. 1758, Hyderabad, India, 2023, https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/4/

Chapter 3: Instantiating an XAI-based decision support system and evaluating its effects in the field

Paper VII

Addressing Learners' Heterogeneity in Higher Education: An Explainable AI-based Feedback Artifact for Digital Learning Environments

Felix Haag
University of Bamberg

Sebastian A. Günther
University of Bamberg

Konstantin Hopf
University of Bamberg

Philipp Handschuh
Leibniz Institute for Educational Trajectories

Maria Klose
Leibniz Institute for Educational Trajectories

Thorsten Staake
University of Bamberg

Proceedings of the 18th International Conference on Wirtschaftsinformatik, Paper nr. 330, Paderborn, Germany, 2023, Best Paper Award, https://aisel.aisnet.org/wi2023/74/

Paper VIII

A Feedback Component That Leverages Counterfactual Explanations for Smart Learning Support

Sebastian A. Günther
University of Bamberg

Felix Haag
University of Bamberg

Konstantin Hopf
University of Bamberg

Philipp Handschuh
Leibniz Institute for Educational Trajectories

Maria Klose
Leibniz Institute for Educational Trajectories

Thorsten Staake
University of Bamberg

In: Digitale Kulturen der Lehre entwickeln – Rahmenbedingungen, Konzepte und Werkzeuge. Edited by L. Mrohs, J. Franz, D. Herrmann, K. Lindner, and T. Staake. Perspektiven der Hochschuldidaktik. Wiesbaden, Germany: Springer VS, 2023, https://link.springer.com/book/9783658433789

Paper IX

Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course

Felix Haag
University of Bamberg

Sebastian A. Günther
University of Bamberg

Philipp Handschuh
Leibniz Institute for Educational Trajectories

Maria Klose
Leibniz Institute for Educational Trajectories

Konstantin Hopf
University of Bamberg

Thorsten Staake
University of Bamberg

Submitted to the European Journal of Information Systems (author's version provided below)

Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course

Abstract

Feedback interventions are a cornerstone of classroom teaching. While prior studies have demonstrated the potential of system-generated feedback in online learning, such interventions often fail to address individual differences in learners' self-regulated learning (SRL), including their motivation, metacognition, and learning behaviors. Consequently, these interventions may disadvantage learners who struggle to plan, monitor, and evaluate their learning. We therefore examine whether SRL theory can inform system-generated feedback that accounts for learner heterogeneity. To this end, we employ Machine Learning (ML) and Explainable Artificial Intelligence to adapt to learners' individual needs. Using over 800,000 platform log events from N=189 university students, we model the relationship between learners' characteristics, platform use, and academic success. Counterfactual Explanations (CEs) then generate actionable feedback tailored to individual learning behavior. An implementation in a university course as a semester-long randomized controlled trial (N=82) shows improved platform engagement and exam performance. Our work makes three contributions: First, we demonstrate that ML captures learner heterogeneity, while CEs translate these insights into personalized instructions. Second, we show that SRL-informed feedback produces large effects on students' academic outcomes and platform use. Third, we highlight the value of adaptive feedback, as students benefit irrespective of their metacognitive skills and procrastination behavior.

Keywords: Higher education; self-regulated learning; learner heterogeneity; explainable artificial intelligence; counterfactual explanations

1 Introduction

Self-Regulated Learning (SRL) theory posits that learners differ in their motivations, metacognitions, and behaviors (Credé and Phillips, 2011; Pintrich, 2000). Acknowledging these inherent differences, teachers aim to tailor their support to the needs of each

individual learner. For instance, a student struggling with time management may receive guidance on setting realistic goals and organizing their study schedule, while a student skilled at memorization might be encouraged to practice deeper critical thinking and transfer exercises. When a teacher opts for replacing this personalized feedback strategy with a more standardized approach, the learning outcomes tend to decrease (Hattie, 2008).

Despite knowing that personalized feedback—considering the individual background and competences of a learner—is more effective than a generalized one (Santhanam et al., 2008), digital learning environments often support only a single learning path, providing standardized instructions to all learners, such as which materials to review or reminders of upcoming deadlines. While recent research shows that some learners benefit from such *non-personalized* support, others struggle to take full advantage, as these learning interventions frequently fail to account for learner-specific needs: Multiple experimental studies highlight considerable disparities in the effectiveness of digital learning interventions (Leung et al., 2023; Li et al., 2021; Mousavi et al., 2025). This is in line with SRL theory, as for example, an intervention leveraging learning goals primarily benefited students with a high prior learning performance and less social activity (Mousavi et al., 2025); an experiment with deadline reminders suggests that this type of feedback negatively affects the performance of students with a low number of courses (Huang et al., 2021). Such interventions disproportionately disadvantage students with specific needs (e.g., with low prior learning performance), thereby exacerbating educational inequalities (Gottschalk and Weise, 2023). This is especially relevant as the increasing shift of learning to digital platforms seems to result in weaker students falling further behind, particularly those from underrepresented or disadvantaged backgrounds (Aucejo et al., 2020). Such a development undermines the sustainable development goals for equitable education (Gottschalk and Weise, 2023; United Nations Department of Economic and Social Affairs, 2024).

Research in Information Systems (IS) has acknowledged that interventions could exacerbate educational inequality, and studies frequently refer to SRL theory to explain the heterogeneous effects observed in digital learning environments (Leung et al., 2023; Mousavi et al., 2025). Yet, these studies often fail to identify underlying causes of heterogeneous outcomes, one of which could be the lack of support for multiple learning paths. Moving from non-personalized to personalized feedback approaches raises the question of how IS scholars can provide support that reflects the individual needs and thus the heterogeneity of learners. As we did not find a satisfactory solution in the literature to address this problem, we explored several approaches for developing more adaptive interventions and tested a promising one that is informed by SRL theory and combines two methods. The first, Machine Learning (ML), that models the behavior of individuals on learning platforms, allowing for a deeper connection between students' learning paths and

their subsequent academic success (Gray and Perkins, 2019; Mubarak et al., 2022). The second, Explainable Artificial Intelligence (XAI), builds on the ML models by employing Counterfactual Explanations (CEs) to generate alternative scenarios and suggest how learning behavior can be improved (Fernández-Loría et al., 2022; Wachter et al., 2018).

We conducted two field studies to evaluate this approach, explicitly considering SRL theory in the provision of learning support. In the first study, we explore the extent to which ML, combined with CEs, can address learner heterogeneity in providing support. Specifically, we technically validated a feedback component based on the association between behavioral patterns and learning outcome. Our approach draws on behavioral data recorded by a typical digital learning environment, with 806,093 log events from two blended learning courses that took place in 2020 and 2021 ($N = 189$ students). In the second study, we embedded this feedback into a digital learning environment in the field and assessed its behavioral effects through a randomized controlled trial conducted within a bachelor-level course in 2022 ($N = 82$ students). Our findings reveal that the generated feedback significantly increases students' learning time and course performance. Furthermore, because the feedback effectively addressed the heterogeneity of learners, we observed consistent desirable effects across different levels of SRL-related factors, meaning that the approach helped students independent of their metacognitive skills and procrastination behavior.

Our studies show that SRL theory informs the development of digital learning environments on how to tackle pressing challenges, such as achieving inclusive and equitable education, by incorporating interventions that consider learner heterogeneity and support multiple learning paths. Specifically, our work makes the following three contributions: First, we demonstrate that ML effectively captures learner heterogeneity, while CEs extrapolate this heterogeneity into personalized feedback. Second, we show that feedback informed by SRL theory produces large effects on both students' platform use and their subsequent exam performance. Third, we advance the discourse on adaptive feedback as our results indicate that students benefit from the feedback, irrespective of their SRL-related factors. Overall, these insights contribute to IS research by (a) emphasizing the importance of addressing learner heterogeneity in digital feedback interventions, (b) demonstrating the value of SRL theory as a framework for guiding the implementation of such interventions, and (c) highlighting the feasibility and effectiveness of ML and XAI-based approaches for providing scalable, personalized feedback.

2 Background and Hypotheses

Before we present our hypotheses, we briefly introduce SRL theory, related work in IS, and explain how ML and XAI could enable the reflection of learner heterogeneity.

2.1 Self-Regulated Learning (SRL)

SRL theory is one of the most influential conceptual frameworks in contemporary educational psychology, offering a comprehensive understanding of how learners regulate their own learning (Panadero, 2017). The ability to self-regulate one's learning (i.e., to plan, monitor, and control the learning process) becomes even more important in higher education, as students are faced with greater autonomy in learning than in secondary education (Vosniadou, 2020). As higher education is constantly shifting from face-to-face to online environments, which are usually characterized by even lower levels of support and guidance, the demands on students in terms of SRL are further intensified (Kizilcec et al., 2017). Consequently, high levels of self-regulation skills (i.e., students who are active in terms of their learning behavior and SRL-related use of strategies) are positively related to academic success (Credé and Phillips, 2011).

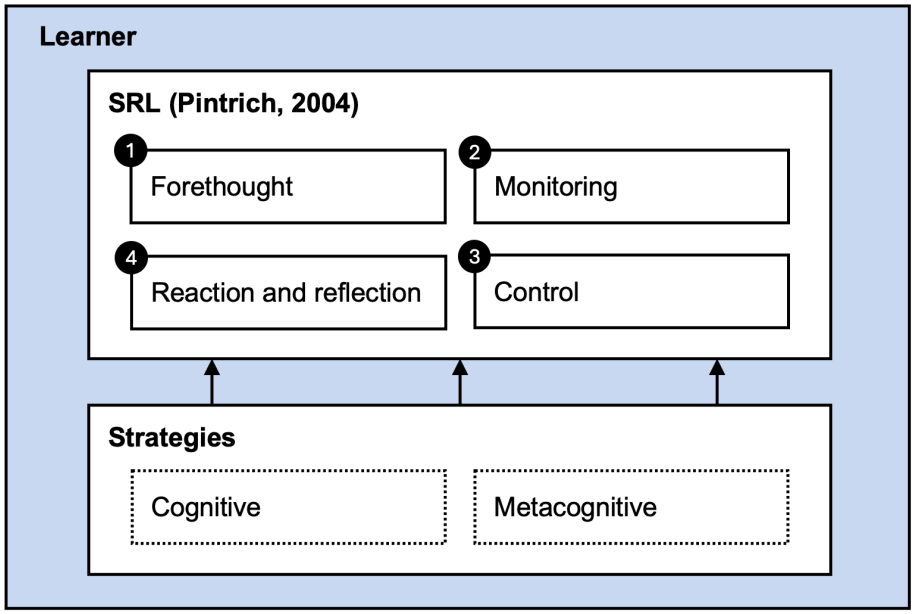


Figure 1: SRL phases and the influence of learner characteristics

In Pintrich's SRL model (Pintrich, 2000; Pintrich, 2004), the learning process can be described by four phases within which learners can decide to apply SRL strategies (see Figure 1). Learners engage in the planning of their activities (phase 1: forethought phase) before they monitor aspects concerning their learning plans (phase 2: monitoring). Subsequently (phase 3: control), they regulate themselves or their environment, for example, by adapting specific learning strategies. In the last phase (phase 4: reaction and reflection), learners reflect about their progress and make judgments about chosen learning strategies (e.g., whether taking notes, self-quizzing, or reviewing material in chunks is helping them understand and retain the content). The effective use of SRL strategies depends on the learner's level of SRL capabilities. For instance, learners with low levels of metacognition might fail to organize their learning in phase 1, while those with higher

levels might activate prior knowledge to identify knowledge gaps for a specific topic that they plan to fill.

SRL skills play a fundamental role in how students manage to cope with their learning. Indeed, these skills are predictors for self-regulation failures, such as procrastination (Pintrich et al., 1993; Steel, 2007). Academic procrastination, in particular—referring to the delay of study-related tasks like exam preparation or fulfilling attendance obligations (Steel and Klingsieck, 2016)—is a frequently highlighted phenomenon in the literature (Pintrich, 2000; Wolters, 2003). A central aspect of procrastination is postponing (i.e., intentionally delaying) a task. This often results in rushing to finish just before a deadline or failing to complete the task within the required time frame (Schraw et al., 2007; Wolters, 2003). As with the use of SRL strategies, learners differ greatly in their tendency to procrastinate: While some learners succeed in completing tasks on time without delay, others struggle and face serious consequences for their academic success (Tuckman, 2005).

Knowing of learner differences in SRL-related factors (e.g., procrastination and metacognitive skills) is fundamental for implementing effective teaching in classroom settings. Similarly, digital learning environments face the challenge of addressing the heterogeneity among learners. Scholars therefore have to acknowledge these differences by trying to implement new feedback approaches that provide personalized support that appropriately fits the learners' individual needs.

2.2 Information Systems and Digital Interventions for Learner Support

The IS discipline has long explored digital interventions that address problems arising from a lack of selected SRL skills (Gupta and Bostrom, 2009). These interventions range from simple prompts, such as deadline reminders, to more sophisticated approaches, including gamification (Leung et al., 2023; Santhanam et al., 2016) and advanced user interfaces aimed at enhancing learners' reasoning skills (Wambsganss et al., 2025).

Existing research on digital behavioral interventions in the IS of digital learning can be grouped into three main areas (see Table 1). First, studies that focus solely on analyzing the overall treatment effect *without conducting subgroup analyses*. While these studies focus on selected SRL-related problems, it remains unclear whether the interventions address the heterogeneity of the learners if they support only a certain subgroup (Leeds et al., 2013; Li and Zhang, 2016; Piccoli et al., 2020). Second, some studies identify which subgroups benefit but still offer 'non-personalized' interventions that use, for example, standardized gamification badges (see, e.g., Leung et al., 2023) that do not adjust their content frame to learners. Consequently, such interventions *do not address learner heterogeneity* and often only benefit specific groups, such as those with high platform loyalty, self-efficacy, or goal orientation. Finally, one study showcases an

Table 1: Related studies examining digital learning interventions

Reference	Mode	Sample size	Intervention	Dependent variable	Effect(s) dependent on
I. No subgroup analyses					
Leeds et al. (2013)	Online course	162	11 learning recommendations	Retention rate	Not examined
Santhanam et al. (2016)	Digital learning (laboratory exp.)	144	Gamification	Computer learning self-efficacy, learning outcomes, enjoyment, focused immersion, and temporal dissociation	Not examined
Li and Zhang (2016)	Massive Open Online Course (MOOC)	570	Monetary incentives, social comparison	Course score	Not examined
Piccoli et al. (2020)	Blended learning	27, not stated, and 310	Performance feedback	Completion rates (of assignments and their chunks)	Not examined
II. Intervention does not address learner heterogeneity					
Gupta and Bostrom (2013)	Digital learning	435	4 different training modes	Learning outcomes, satisfaction, cognitive knowledge	Faithfulness of learning platform, level of collaboration
Huang et al. (2021)	MOOC	7,844	5 treatment messages	Assignment timeliness and completion rate	Students' active course load, tenure on the learning, prior grade, education level
Leung et al. (2023)	MOOC	838	Gamification	SRL engagement and learning outcomes	Performance-approach, performance avoidance, and mastery goal orientation
Mousavi et al. (2025)	MOOC and a smaller online course	1,451 and 581	Various goal conditions	Goals of learning, performance-prove, and performance avoidance	Student's prior performance, social activity
III. Intervention does address learner heterogeneity					
Wambsganss et al. (2025)	Digital learning (laboratory exp.)	54, 142, and 124	Writing support	Writing performance (objective and subjective quality of argumentation)	Heterogeneity tests showed no significant differences in argumentation by prior learner expertise.

intervention *that addresses learner heterogeneity* by tailoring feedback to learners' needs: Wambsganss et al. (2025) introduced an ML-based approach that adapts to individual learner differences, demonstrating in a laboratory setting that its effectiveness is not moderated by prior argumentation skills. The study's intervention uses natural language processing and is limited to supporting writing skills.

To sum up, previous studies have either neglected learner heterogeneity (e.g., Leung et al., 2023) or focused on specific tasks within controlled environments (Wambsganss et al., 2025). Despite recent technological advances, such as those enabled by ML, it remains unclear whether current interventions can effectively address learner heterogeneity for general learning tasks and thus mitigate the inequalities fostered by 'non-personalized' approaches.

2.3 Machine Learning (ML) and Explainable Artificial Intelligence (XAI) for Learner Support

Recent studies in the field of learning analytics that employ ML mainly focus on predicting students' learning outcomes (Gray and Perkins, 2019), early drop-outs from courses (Mubarak et al., 2022), and the identification of learners with other learning difficulties (Delen et al., 2024; Hoffait and Schyns, 2017). Such predictions are primarily informative for instructors, who can intervene and proactively support students at risk. Similarly, learning-analytics systems use platform engagement data to provide instructors with actionable information for monitoring students and intervening where needed (Nguyen et al., 2021).

Yet, the patterns that ML detects in data are also interesting for developing individual learning advice (Delen et al., 2024). This is where XAI—a recent stream in ML research that revolves around the exposure of ML-detected patterns in data—comes into play. XAI methods have the capability to make the patterns that ML detects in the data transparent to users. In controlled settings, XAI has been shown to produce desirable effects such as improved task performance and learning outcomes (Förster et al., 2025). While many XAI approaches revolve around the influence of predictor variables (Lundberg et al., 2020), leading to merely descriptive information, a promising approach for prescriptive insights is CEs. The concept of CEs involves causal answers to "what-if?" questions for a given phenomenon, e.g., "if a learner consumes a specific set of learning material, she would achieve a better grade" (Fernández-Loría et al., 2022; Miller, 2019). In the context of ML, CEs provide explanations in the form of required changes for a desired model outcome (e.g., a higher predicted exam grade). In other words, a counterfactual is the smallest change in the input features that alters the prediction to a (predefined) output (Mothilal et al., 2020). Such explanations might translate into specific real-world cause-effect relationships, e.g., "If I had spent more time on the learning platform in this

section, I would have had a better grade in the exam".

CE-based feedback takes different learning paths and, thus, the heterogeneity of learners into account. This may benefit higher education learning in several ways. First, such explanations can inform individuals, in the form of feedback, on how to achieve a desired learning outcome (e.g., more exam points). Second, exposing the behaviors of strong learners to potentially struggling learners has already been found to be effective in prior SRL studies (Davis et al., 2017). As ML uses data on past learning behavior that has led to high and low learning success, the approach could suggest learning approaches from strong learners to weaker learners and thus may reduce heterogeneity in the application of metacognitive learning strategies. Finally, providing guidance on a personal level that is automated does not require the additional involvement of instructors and may even be advantageous, as learners have been found to learn better from critical feedback delivered by Artificial Intelligence (AI) than by humans (Turel, 2025).

Current implementations of ML and related technologies for guiding digital learning often fail to account for variations in learners' SRL skills. Thus, investigating feedback that utilizes ML and CEs is of particular interest to the IS field, given their potential to model the heterogeneity of learners and translate these patterns into actionable instructions.

2.4 Hypotheses Development

The diversity in learners' backgrounds (e.g., prior knowledge or study program) and the heterogeneity in their SRL skills and strategies (Credé and Phillips, 2011; Pintrich, 2000) are reflected in the nonlinear learning behavior captured in learning platform log data. Capturing such heterogeneity poses a notable challenge. Advances in learning analytics indicate that ML can leverage diverse behavioral patterns recorded in digital learning environments (Delen et al., 2024; Gray and Perkins, 2019; Mubarak et al., 2022) to generate feedback tailored to individual needs. Yet, merely identifying heterogeneity does not necessarily improve learning outcomes—the critical challenge lies in translating these insights into actionable and personally meaningful instructions.

Drawing on SRL theory, we argue that effective feedback should translate complex, data-based insights into personalized guidance. While ML can model behavioral heterogeneity, XAI—in particular, CEs—can transform these patterns into meaningful feedback. CEs estimate how specific changes in learning behavior (e.g., increasing quiz engagement) are likely to affect learning outcomes, thereby providing actionable and personalized instructions aligned with each learner's progress. In this way, CE-based feedback could enable data-driven personalization that directly addresses the learner heterogeneity emphasized by SRL theory.

Building on this rationale, we posit that CE-based feedback should positively influ-

ence both engagement and performance. Specifically, we expect CEs to increase learner outcomes in three ways: First, CEs can increase learning time on the platform. This extended learning time may result from greater engagement with quizzes, videos, and slides. Second, increased learner engagement with the proposed feedback content may lead to a higher likelihood of course completion and taking the final exam, as learners might perceive 'sunk costs' when considering dropping out (Coleman, 2010). Additionally, higher completion rates may also be driven by an improved subjective sense of the feedback's positive impact or a more accurate assessment of their own performance. Third, the behavioral patterns identified through CE-based feedback may provide genuine value, helping learners achieve higher exam performance. Accordingly, we hypothesize:

Hypothesis 1 *Providing learners with counterfactual explanation-based feedback during online learning* ***(a)*** *increases online learning time,* ***(b)*** *increases exam-taking rate, and* ***(c)*** *increases exam performance.*

The prior hypotheses raise an important question: Does such feedback benefit all learners, or only those with a specific level of SRL? While low metacognition and high procrastination are often seen as barriers to academic success (Everson and Tobias, 1998; Steel, 2007), CE-based feedback may reduce these effects by offering tailored guidance that addresses individual needs. For instance, learners with lower metacognition could receive basic instructions (e.g., "You should spend more time on the platform to increase success"), while higher metacognition learners get more focused content (e.g., "Try to solve quiz 7 again to check your knowledge"). Similarly, the feedback can precisely target next learning steps for learners by taking their prior behavior into account. For example, learners who procrastinate during the semester and delay their learning until just before the exam should receive different suggestions before the final examination than those who have been consistently active. Therefore, the effects of CE-based feedback are expected to operate independently of SRL-related factors such as metacognition or procrastination, as the feedback inherently accounts for individual learning paths. Following this line of reasoning, we hypothesize:

Hypothesis 2 *The effects of counterfactual explanation-based feedback on exam performance are independent of important self-regulated learning-related factors (i.e., metacognition and procrastination).*

3 Methods

In this section, we begin with describing the university course, which serves as the foundation for both field studies. Next, we present our ML approach for modeling individual

learning paths, followed by a description of how we implemented XAI and CEs to generate heterogeneity-aware feedback. Finally, we outline the randomized controlled trial designed to test the effectiveness of the feedback.

3.1 Course Description

We ran our studies in an IT project management Bachelor's level course, which is mandatory for the IS program at our institution. It covers topics such as project planning, agile software development (e.g., Scrum), and key performance indicators. The course is delivered in a blended learning format: Weekly lectures were provided as on-demand videos through the accompanying digital online learning platform, implemented using Open edX. Additionally, four sessions, including the introductory lecture and Q&A, were held in person. Tutorials took place once a week in a traditional classroom setting. We used the digital online learning platform to distribute all course materials, such as tutorial slides and voluntary quizzes, to learners. To pass the course, learners have to successfully participate in an exam at the end of the semester, which covers topics from the lectures and tutorials. Students have the opportunity to retake the exam until they pass (this is possible due to study progress controls in the program[1]). The lecture period covers 14 weeks, where the exam is scheduled by the examination office between one and five weeks after this period.

3.2 Study 1

To model learner heterogeneity, we collected data from two runs of the course. We used ML to model individual learning paths and used XAI and CEs to generate feedback.

3.2.1 Data collection and preparation.

We collected several types of data on the online learning platform (see Table 2). These data include learning time (e.g., how much time a learner spends in a specific learning section) and platform activity events (e.g., when a particular video has been viewed entirely), recorded in a database using a JavaScript plugin and logging functions (e.g., platform activities, files downloaded, and videos watched). To further account for learners' heterogeneity, we also collected data on individuals' characteristics (e.g., the program of study) through a built-in survey. In summary, the learning data for ML model instantiation comprise 189 learners resulting in 806,093 platform activity events (e.g., videos watched) and 64,207 learning time events (e.g., time spent in a specific section). We finally harmonized the resulting events and learner characteristics from both course runs, extracted features that can be fed into ML algorithms. We implemented an expanding

[1] In the IS program of the institution, students have to earn 50 credits by the end of the fourth semester.

time window approach, which means that with each week of the semester, the number of data points and features grows as additional learning materials and behavioral data are continuously added. This data reflects students' learning behavior over time.

Table 2: Overview of course data used for model training

Semester	Summer 2020	Summer 2021
Students	95	94
Female	28	29
Male	63	61
Other	0	1
Not reported	4	3
Exam points mean (SD)	53.58 (14.22)	52.59 (15.60)
Data collection period	Apr 20–Aug 12, 2020	Apr 12–Jul 19, 2021
Learning time events	34,048	30,159
Platform activity events	418,356	387,737

3.2.2 ML and Counterfactual Explanation (CE) Modeling.

To model learners' behavior, we instantiated ML models that predict learners' exam score ($0 - 90$ points), as a proxy for course success. We trained separate models for each week using an expanding-window approach: the model for week t included all learner data from weeks $1 \ldots t$ to predict exam scores, thereby reflecting the sequential nature of learning. This approach ensured that predictions of the exam scores leveraged the accumulated behavioral data of learners rather than only isolated weekly snapshots. For the prediction of the exam scores, we used ML algorithms that stem from different categories and are known to perform well for predicting academic performance (e.g., Miguéis et al., 2018), namely Extreme Gradient Boosting, CatBoost, Random Forest, and Support Vector Regression. To assess the resulting models in an unbiased manner, our underlying procedure performed for each week in the semester a parameter tuning using a 10-fold cross-validation grid search with 80 % of the data and then selected the model yielding the lowest prediction error according to Root Mean Square Error (RMSE) on the remaining 20 % (i.e., the test set). This approach follows the assumption that the more accurately a model recognizes domain concepts (which are approximated through predictive performance), the higher the quality of the resulting data-driven explanations and vice versa (Štrumbelj et al., 2009). Finally, we fit models on the entire dataset (i.e., all prior course data for a specific week) that serve as the foundation for counterfactual modeling and, ultimately, feedback generation.

The counterfactual modeling relied on the approach outlined by Mothilal et al. (2020) to uncover the pattern the models detect in the data. For this purpose, we used the model trained on historical data up to week t and applied it to the current course's

behavioral data up to the same week in order to generate counterfactual feedback for week $t+1$ (e.g., a model trained on historical data up to week 8 was applied to the week 1–8 data of the current cohort to produce feedback shown in week 9). Using this approach, the feedback's underlying algorithm calculated the largest possible shift towards improved academic performance (i.e., more points in the exam) and determined the associated behavioral changes in the digital learning platform (i.e., changes in the feature values). The logic selected the three feature changes with the highest impact on the exam points, and then stored all changes for each learner in an operational database. Finally, the feedback component translated the feature changes into actionable instructions.

3.2.3 Feedback Intervention.

The design of our feedback relies on requirements identified from established knowledge in SRL theory and the XAI literature to adequately address learner heterogeneity. From a technical perspective, we implemented a feedback component that can be displayed on the course home page of the edX learning environment, enabling learners to receive personalized instructional feedback (Figure 2). The feedback is structured into three key parts: First, the introductory text, that points learners to the tailoring to the personal past learning behavior on the platform. The feedback thereby highlights the counterfactual outcome for learners by explaining that additional implementation of these instructions can lead to improved exam performance. Because the information that the feedback is XAI-based could potentially distract the learners from the feedback content, we decided to exclude this from the introductory text. Second, actionable instructions were purposefully restricted to three items, as we argue that adding more might induce reactance due to a substantial increase in additional workload. Third, check boxes are positioned next to each instruction to help learners track their progress of feedback implementation.

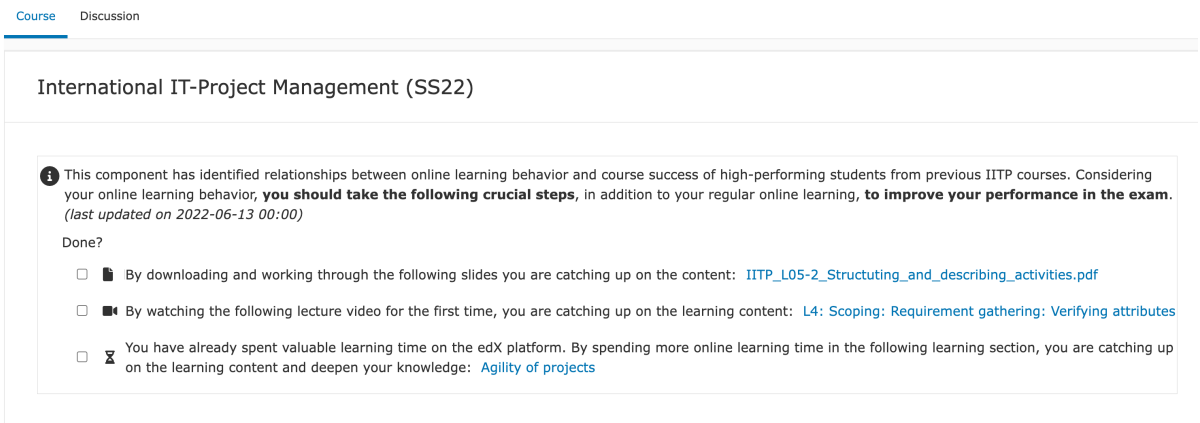


Figure 2: Exemplary feedback intervention

3.3 Study 2

To experimentally evaluate the effects of the feedback, we conducted a randomized controlled trial in a semester following the training data collection.

3.3.1 Experimental Design.

Our experiment follows a typical difference-in-differences experimental design, which is displayed in Figure 3. We began with a six-week baseline phase to observe online learning behavior without any experimental manipulation. After this period, we divided the students of a new bachelor course in 2022 into two groups: a control group (N = 41) and a treatment group (N = 41). To do so, we employed a stratified randomization based on the students' study program. From week seven, participants in the treatment group could see the feedback in the digital learning environment, whereas those in the control group did not receive any intervention. The lecture period covered 14 weeks, with the exam scheduled five weeks later[2]. Feedback was updated weekly throughout the intervention phase (i.e., from week seven to nineteen) around midnight from Sunday to Monday.

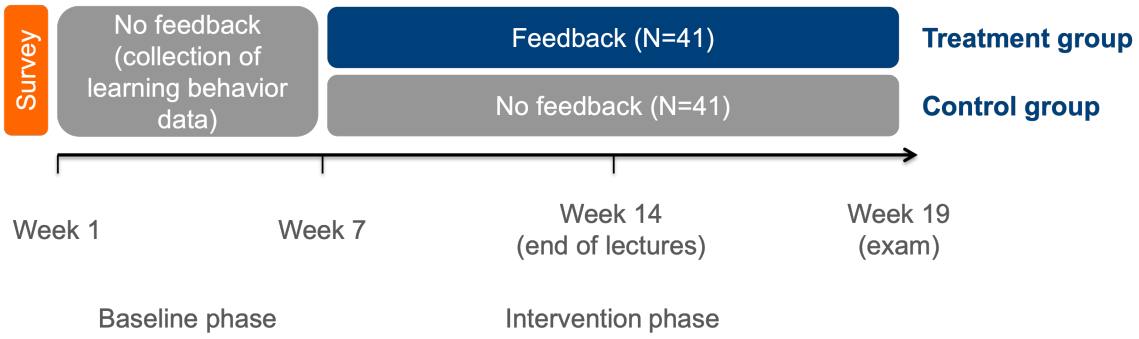


Figure 3: Difference-in-differences experimental design

3.3.2 Sampling and Data Collection.

During course registration, we asked learners whether they were willing to participate in our experimental study. If so, we asked questions about their sociodemographics, SRL strategies, and procrastination behavior. We measured metacognition by assessing the use of metacognitive strategies, adapting the 12 items introduced by Pintrich et al. (1991) (rated on a 7-point Likert scale). For academic procrastination, we adapted the short scale by Yockey (2016) that builds on the work of McCloskey (2011). The scale comprises 5 items and measures learners' academic procrastination on a 5-point Likert scale. As in study 1, we tracked data on learner behavior on the platform and student characteristics.

[2]The experiment started on Apr 25, 2022, and the exam took place on Aug 31, 2022.

3.3.3 Pre-processing Procedures.

For a plausible estimation of behavioral response to the feedback, we had to pre-process the data of the course. We excluded 13 out of the 82 accounts of the learning platform, as the users never went online again after the baseline phase. Finally, we excluded three students who had not registered for the exam (i.e., they decided in the baseline phase and at the latest two months before the exam not to show up on the final examination date without consequences), as we wanted to focus only on learners who shared the goal of passing the course. Overall, this resulted in a dataset of 66 learners.

3.3.4 Descriptive Statistics and Balance Check.

Before exploring behavioral effects, we performed a randomization check on the pre-processed data to verify that the randomization procedure was successful (Table 3). The first column describes the variables on which we have performed the randomization checks. Column 2 presents the corresponding mean value of the sample with standard deviation for the respective variable. The columns 3 and 4 display these values for the control group and the treatment group. The fifth column shows the F test statistics with p-values in parentheses. The table indicates that our groups are balanced in key characteristics.

Among the 66 learners enrolled in the course, 27 identified as female and 38 as male (one participant did not report their gender). The average age was $M = 23.0$ years ($SD = 4.79$), based on responses from 62 students (four were missing).

3.3.5 Propensity Score Matching (PSM).

Following Leung et al. (2023), we applied PSM to further mitigate the risk of sampling bias and to increase the validity of comparisons between groups for the analyses on SRL-related factors. The approach allows matching each learner in the control group with a study participant in the treatment group that is similar in terms of their SRL-related characteristics and online learning behavior. We used optimal matching with a 1:1 ratio and without replacement (i.e., the matching uses each participant in the control group only once) to avoid overrepresentation of specific control group participants and ensure appropriate matching for the given sample. The propensity scores were estimated through logistic regression, with the treatment indicator regressed on the SRL-related factors and platform use. Our PSM analysis estimates the average treatment effect on the treated using g-computation for the pre-specified outcome model and standard errors and confidence intervals using cluster-robust standard errors with pair membership as the clustering variable (see supplementary material for more details).

Table 3: Randomization check

Variable	Full sample	Control group	Treatment group	F-Statistics (p value)
Average online learning time per week [in min]	29.88	31.22	28.61	0.153
	(26.87)	(26.84)	(27.24)	(0.697)
Proportion of women	0.41	0.41	0.41	0.002
	(0.50)	(0.50)	(0.50)	(0.964)
Self-efficacy[a]	4.82	4.79	4.85	0.071
	(0.85)	(0.82)	(0.89)	(0.791)
Metacognition[a]	4.58	4.65	4.53	0.322
	(0.83)	(0.70)	(0.93)	(0.573)
Cognition[a]	4.49	4.58	4.42	0.424
	(0.96)	(0.87)	(1.03)	(0.518)
Management of internal resources[a]	5.04	5.11	4.99	0.229
	(0.92)	(0.74)	(1.06)	(0.634)
Procrastination[b]	2.53	2.44	2.61	0.465
	(0.92)	(0.80)	(1.02)	(0.498)
Test anxiety[a]	3.62	3.78	3.48	0.974
	(1.18)	(1.23)	(1.14)	(0.328)
Conscientiousness[b]	3.23	3.24	3.22	0.020
	(0.62)	(0.44)	(0.76)	(0.889)
Retention rate[c]	0.80	0.78	0.83	0.304
	(0.40)	(0.42)	(0.38)	(0.583)
N	66	32	34	–

Notes: [a] Scored using a seven-point Likert scale; [b] Scored using a five-point Likert scale; [c] Resulting percentage after filtering from the original randomization sample that did not drop out of the study.

4 Empirical Results

We first examine whether ML and CEs reflect learner heterogeneity, which serves as a technical foundation for the subsequent study. Then, we analyze learners' behavioral responses to the CE-enabled feedback (Hypotheses 1a–c), also in light of certain SRL-related factors (Hypothesis 2).

4.1 Study 1

4.1.1 ML to Model Learner Heterogeneity.

To evaluate the reflection of learner heterogeneity through ML, we follow Štrumbelj et al. (2009) by measuring predictive performance using the widely recognized regression error metrics Mean Absolute Error (MAE) and RMSE (see Hastie et al., 2009 for details on these measures). Our results show that error rates decrease as weekly data accumulates

for all ML models from week four onwards, with performance overall stabilizing in the subsequent weeks. Specifically, across all weeks and best models, we find the following error metrics: MAE = 9.63 and RMSE = 12.20.

To further validate the effectiveness of the ML models, we compare their residuals with those from a naïve estimator that uses the mean or median of the training set. The mean estimator yields an MAE of 11.389 and an RMSE of 14.178, while the median estimator resulted in an MAE of 10.802 and an RMSE of 14.171. The ML models outperform naïve methods in a potential baseline and intervention phase. We confirm this with paired two-sided t-tests comparing the RMSE on the test set (i.e., best-performing ML model for the respective week vs. mean estimator). The tests yield a significant reduction in RMSE for the baseline phase ($t(5) = -6.820$, $p = 0.001$, $d = 2.692$) and the intervention phase[3] ($t(9) = -12.617$, $p < 0.001$, $d = 4.033$). In addition, we observe a significant decrease ($p = 0.009$) in residuals by 15.16 % through the best-performing ML model in comparison to the mean estimator across all weeks and users in the test set[4]. These results suggest that the ML models are effective in capturing meaningful patterns in students' learning behavior.

4.1.2 XAI to Reflect Learner Heterogeneity.

In addition, we need to confirm that our technical approach successfully produced personalized feedback content. To achieve this, we analyzed the diversity of the learning instructions by counting the number of unique actionable features (e.g., learning time in a specific section) across study participants. A chi-squared test indicates a statistically significant difference in instructions between participants ($\chi^2 = 6154.8$, $p < 0.001$ using the method of Hope, 1968). In addition, the Gini index, measuring the spread in instructions, is 0.38, indicating that the instructions provided are distinct and only a few receive uniform feedback content.

4.2 Study 2

4.2.1 Students' Implementation of the Feedback.

First, we check whether the course participants have actually increased their online learning time in response to feedback. To this end, we estimate the effects of the feedback on the online learning time based on the following equation, using ordinary least squares:

$$y_{it} = \alpha_i + IN_{it} \times (\beta_0 + \beta_1 T_i^{Feedback}) + \varepsilon_{it} \quad (1)$$

[3]In comparison to study 2, the semesters in summer 2020 and summer 2021 had fewer weeks until the exam.

[4]For this analysis, we ran a regression that estimates the absolute error while clustering the standard errors across test set users.

where y_{it} is the online learning time of participant i in week t. We include an individual fixed effects parameter α_i for each participant to eliminate the variance resulting from fixed differences across the participants. The binary variable IN_{it} is all 0 during the baseline phase and takes the value of 1 in the intervention phase. $T_i^{Feedback}$ is a binary variable that indicates whether a participant i belongs to the control group ($T_i^{Feedback} = 0$) or the treatment group ($T_i^{Feedback} = 1$). We cluster the standard errors on the participant level and show the effects of the feedback on online learning time in Table 4.

Table 4: Impact on online learning time

Independent variable	Online learning time [in min per week]
	Model 1
Intervention phase	-14.45^{***}
	(4.37)
Intervention phase × Treatment	$+12.49^{**}$
	(5.95)

Notes: *, **, and *** indicate significance at the 10 %, 5 %, and 1 % levels, respectively.

Our base analysis of the effect on learning time yields two main findings. First, course participants were overall less active on the platform during the intervention phase. Second, feedback counteracted this trend. The control group reduced their learning time by 14.45 minutes per week during the intervention phase, spending 15.48 minutes on the platform. In contrast, the treatment group spent 27.97 minutes per week [5]—an increase per week of 12.49 minutes (or 80.69 %) in comparison to the control group in the intervention phase ($p = 0.046$). In summary, the treatment effect amounts to 2.71 hours of additional online learning time. We therefore find support for Hypothesis 1a.

Next, for in-depth analysis, we check whether the participants have indeed implemented the learning instructions based on the feedback (in comparison to the regular learning behavior of the control group). To this end, we take all learning instructions that have been displayed to the course participants and analyze whether the participants have actually followed them (e.g., actually solved a quiz when the feedback contained this suggestion). In addition, we control for the aspect that course participants in the treatment group might not have implemented instructions because they have paid attention to the feedback element, but rather because they wanted to perform the respective action independently of any external manipulation. To this end, for each instruction, we have determined the number of participants in the control group who have also implemented the underlying learning action. In doing so, we compute for each treatment group participant

[5]The treatment group reduced learning time during the intervention phase by only 1.96 minutes compared to the baseline phase. Also, this small difference is not statistically significant ($p = 0.628$).

whether the percentage of implemented instructions exceeds the percentage of those from the control group. Indeed, a corresponding t-test indicates that participants in the treatment group have intentionally implemented the instructions ($p = 0.025$). Both results, namely increased online learning time and learning material consumption[6], highlight that the feedback increased online activity.

4.2.2 Impact on Course Performance.

To derive the impact of the feedback on course performance, we check whether the feedback has materialized into a higher exam-taking rate and exam performance. A logistic regression model described by the following equation estimates the effect on learners' exam-taking rate:

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \alpha + \beta_0 T_i^{Feedback} \tag{2}$$

where y_i is a binary indicator that determines whether a participant i has shown up for the course's final exam. We display the results of the exam-taking rate in Table 5 (Model 2a). Although we notice a positive trend towards higher participation, the model indicates no statistically significant impact of the feedback on the exam-taking rate. As the effect could also depend on the number of weeks an individual saw the feedback, we also controlled for the number of weeks a participant was online (Model 2b). The point estimates and the standard errors, however, stay relatively similar. We therefore find no support for Hypothesis 1b.

Table 5: Impact on exam-taking rate

Independent variable	Exam-taking rate	
	Model 2a	Model 2b
Constant	+0.938** (0.393)	+0.997** (0.410)
Treatment	+0.820 (0.624)	+0.900 (0.646)
Control for weeks online	✗	✓

Notes: In the second model, we include 'weeks online' as a centered variable to improve the interpretability of the intercept, allowing for easier comparison with the first model, which does not include this variable. *, **, and *** indicate significance at the 10 %, 5 %, and 1 % levels, respectively.

Next, we check how the participants responded to feedback in terms of the exam performance. Based on the following equation, we estimate an ordinary least squares

[6] In additional analyses, we find that participants responded to the feedback by interacting more with videos ($p = 0.081$).

model to estimate the effect of the feedback on the exam score:

$$y_i = \alpha + \beta_0 T_i^{Feedback} + \varepsilon_i \tag{3}$$

where y_i denotes the score in the exam. Participants who registered for the exam but did not take part received 0 points from the examination office. Table 6 displays the impact of the feedback on the exam score. The results indicate that the feedback had a noticeable positive effect on the overall score of the learners ($p = 0.064$, $\eta_p^2 = 0.053$). This also holds true when controlling for the number of weeks online ($p = 0.029$, $\eta_p^2 = 0.073$). Overall, we therefore find evidence for a positive effect on exam scores (Hypothesis 1c).

Table 6: Impact on exam score

Independent variable	Score in exam [0,90]	
	Model 3a	Model 3b
Constant	+33.28***	+33.22***
	(5.01)	(4.27)
Treatment	+13.18*	+13.29**
	(6.98)	(5.95)
Control for weeks online (centered)	✗	✓

Notes: In the second model, we include 'weeks online' as a centered variable to improve the interpretability of the intercept, allowing for easier comparison with the first model, which does not include this variable. *, **, and *** indicate significance at the 10 %, 5 %, and 1 % levels, respectively.

To ensure the robustness of our findings, we conducted an additional analysis to better address the fact that no-shows were treated as having 0 points. For this purpose, we fit a Tweedie regression model using a log-link function to appropriately handle the distribution of the data. Mathematically, we estimate the following relationship and determine the index and dispersion parameters using the profile likelihood approach:

$$\text{Score\_exam}_i \sim \text{Gamma}(\mu_i, \phi) \tag{4}$$

$$\log(\mu_i) = \beta_0 + \beta_1 \text{group}_i + \log(\text{Exposure}_i) \tag{5}$$

where $Exposure_i$ is the percentage of weeks individuals were online in the study. As Table 7 shows, the treatment effect is still marginally statistically significant (p=0.054), demonstrating the robustness of our findings.

Table 7: Robustness check: Impact on exam score with Tweedie regression

Independent variable	Score in exam [0,90]
	Model 4
Constant	$+4.083^{***}$
	(0.139)
Treatment	$+0.357^{*}$
	(0.182)

Notes: *, **, and *** indicate significance at the 10 %, 5 %, and 1 % levels, respectively.

4.2.3 Heterogeneity Analyses in Terms of SRL-related factors.

Next, we investigated whether the treatment effects depend on learners' characteristics. We performed initial matching with the group membership regressed on *procrastination behavior* and *the use of metacognitive strategies* as well as *number of weeks learners accessed the platform* to compute propensity scores (see subsection 3.3.5 Propensity Score Matching).

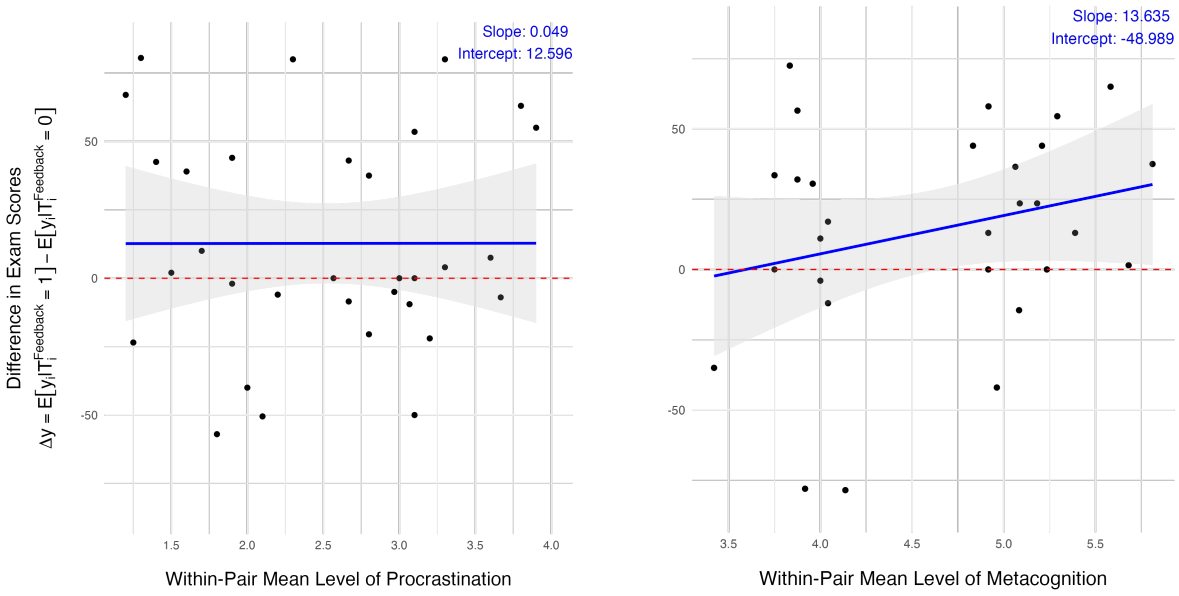
For the estimation of the treatment effect in terms of SRL-related factors and platform use, we adapt Equation 3 by including the respective variables in the outcome model both as standalone terms and as interaction terms with the treatment variable:

$$y_i = \alpha + \beta_0\, T_i^{\text{Feedback}} + \beta_1\, \text{SRL}_i + \beta_2\, W_i^{\text{Online}} + \beta_3\, \text{SRL}_i \times T_i^{\text{Feedback}} + \beta_4\, W_i^{\text{Online}} \times T_i^{\text{Feedback}} + \varepsilon_i \tag{6}$$

where W_i^{Online} denotes the number of weeks a learner accessed the platform and SRL_i represents the specific SRL factor.

Including all SRL-related factors in the outcome model results in a treatment effect of 12.0 exam points ($SE = 6.42$, $p = 0.062$), suggesting that feedback is effective across similar levels of procrastination, the use of metacognitive strategies, and repeated platform activity. To deepen our understanding of this effect, we conducted two separate moderation analyses in relation to procrastination and metacognition by performing exact matching on the respective SRL variable (see the supplementary material for further details on the matching approach for all PSM-based analyses).

First, we analyzed the moderating role of reported procrastination behavior by dichotomizing the variable through a median split and incorporating it into the matching procedure. Figure 4a illustrates the relationship between the within-pair mean procrastination level and the difference in exam scores for each pair (the gray area represents the 95 % confidence interval). The fitted regression line shows a trend close to zero with a positive intercept, indicating an overall positive impact of feedback regardless of procrastination levels. Indeed, a one-sided t-test on the differences in exam score outcomes shows that the effect is significantly greater than zero ($t(31) = 1.81$, $p = 0.040$, $d = 0.321$),



(a) Procrastination (b) Metacognition

Figure 4: Difference in outcomes across SRL-related factors.

which confirms the overall beneficial effect in this matching scenario. To examine subgroup differences, we included the treatment variable, the median split indicator (coded as 1 for "high" procrastination and 0 for "low" procrastination), and their interaction term, $T_i^{\text{Feedback}} \times \text{Proc}_i^{\text{High}}$, in the outcome model. Consistently, we found no significant differences between the subgroups on the treatment effect ($\beta = -1.01$, $SE = 15.0$, $p = 0.946$), indicating that procrastination does not moderate the effect of the feedback.

Second, we investigated the use of metacognitive strategies as a moderator. To this end, we apply the same procedure as for procrastination by conducting a median split on the metacognition variable and performing exact matching based on the resulting groups (i.e., "high" and "low" use of metacognitive strategies). The regression line in Figure 4b suggests that learners with higher metacognition levels tend to benefit more from the treatment. An initial one-sided t-test on the differences in exam score outcomes shows that the overall effect of feedback is positive and significantly greater than zero ($t(28) = 1.97$, $p = 0.030$, $d = 0.365$), indicating a general beneficial impact of feedback across metacognition levels in this matching scenario. As with procrastination, we modeled the outcome using the median split indicator as the SRL variable for moderation analysis. Although the treatment significantly improves exam scores for learners with a high use of metacognitive strategies ($\beta = 22.3$, $SE = 7.34$, $p = 0.002$), the difference in treatment effects between learners with high and low levels is not statistically significant ($\beta = 18.8$, $SE = 14.8$, $p = 0.202$). Hence, we do not find a moderating effect of learners' metacognition and procrastination levels, which supports Hypothesis 2.

As a final step in our heterogeneity analyses, we aim to further understand what drives the treatment effects in relation to SRL. Specifically, we want to determine whether

learners who are rarely online also benefit from the feedback, or if the feedback only demonstrates its full potential for those who frequently see it. To this end, we divided our sample that had registered for the exam into three groups. Specifically, based on the number of weeks the learners accessed the platform, we created three terciles (T1 to T3). We then conducted a regression analysis to estimate the average exam score within each tercile, both with and without feedback. Finally, we compared the treatment subgroups to their respective control groups using the regression parameters (Table 8).

Table 8: Difference in exam score in relation to platform use

Comparison	Difference in exam score
Treatment_{T1} vs. Control_{T1}	$+24.32^{*}$
Treatment_{T2} vs. Control_{T2}	-0.18
Treatment_{T3} vs. Control_{T3}	$+20.68^{*}$

Notes: To account for multiple testing, we adjusted the p-values through the false discovery rate (see Benjamini and Hochberg, 1995). *, **, and *** indicate significance at the 10 %, 5 %, and 1 % levels, respectively.

We see that users with both low and high platform usage benefit particularly from the feedback with respect to their exam performance. One explanation for this is that learners with less online activity (and who show indications of procrastination behavior) can benefit even from minor behavioral changes in their learning. In turn, those with very high platform usage might benefit, as they can implement the feedback as intended.

5 Discussion

Prior studies have mainly employed non-individualized behavioral interventions, where the effectiveness depends on learners' level of SRL skills (see, e.g., Leung et al., 2023) and socio-demographic factors (see, e.g., Li et al., 2021). To overcome this limitation, we developed and tested an XAI-based approach that models the heterogeneity of learners in the form of prior course data and thereby enables personalized guidance in digital learning environments. This work makes several theoretical and practical contributions to IS in higher education, the use of XAI for generating personalized feedback, and behavioral theory-driven design of such IS, which we outline below.

5.1 Summary of Findings

Our real-world experiment with learners who participated in an actual open online learning course leads to the following five findings. First, our ML models reflected learner

heterogeneity and, second, the CE method translated these patterns successfully into actionable instructions for learners. These findings demonstrate the feasibility of reflecting learners' heterogeneity, leveraging ML and XAI. Third, learners who received feedback spent 80.69 % more time on the platform than those in the control group (Hypothesis 1a). We also observed an increased engagement for the treatment group within the online learning platform, such as interacting more with the videos. Learners thereby followed the instructions provided in addition to their own planned actions. This increased engagement resulted in improved academic outcomes, as participants in the treatment group scored on average 13.18 points more in the exam (with a total of 90 points) than the control group (Hypothesis 1c). We also find a trend towards a higher exam-taking rate, though not statistically significant (Hypothesis 1b). This result indicates an increased course engagement and commitment due to learners' prior investment in the course (Coleman, 2010). In general, we observe that our ML- and XAI-based feedback affects academic outcomes as intended.

Fourth, our feedback appears to increase course performance independent of important SRL-related factors, namely procrastination and metacognition (Hypothesis 2). Specifically, we observe a consistent positive impact on learners' exam results, regardless of their level of procrastination. This finding is particularly noteworthy, as procrastination is often viewed as a significant SRL challenge, frequently linked to inefficient learning strategies such as poor time management and the ineffective use of metacognitive strategies (Cerezo et al., 2017; Zhou et al., 2022). This result highlights the effectiveness of tailoring the feedback to individual needs.

Additionally, learners with higher abilities in applying metacognitive strategies show a slight tendency to make greater use of the feedback's beneficial effects. However, we do not find a moderating effect of metacognition but, on average, an increase in the exam score for participants in the treatment group. This tendency may stem from monitoring processes and knowledge of learners' cognition, which are part of metacognition (Pintrich et al., 2000). By frequently using and understanding cognitive strategies, learners may not only develop greater proficiency in applying them, but also enhance their ability to implement feedback effectively. As a result, those with more experience in applying these strategies could make greater use of the feedback provided. A complementary, expertise-based explanation for this phenomenon comes from Kalyuga (2007), who posits that learners with varying levels of competence derive different benefits from personalized instructional systems. Whereas learners at low expertise levels receive meaningful support from relatively basic learning instructions for their progress (e.g., "check this summary for section X"), those at higher levels benefit from nuanced instructions, which means more specific hints that may help to consolidate their learnings and further close potential knowledge gaps (e.g., "review quiz Y"). We attribute this finding to the ability of our approach to model learner heterogeneity, accounting for individual learners' needs. This

result is in stark contrast to other studies that find benefits primarily for certain learners using non-instructional behavioral interventions (e.g., peer information, see Li et al., 2021, or gamification approaches, see Leung et al., 2023). In doing so, our study complements the recent findings of Wambsganss et al. (2025) who observed beneficial effects for students at various levels of expertise in text-writing. However, these effects are specific to supporting students in writing tasks, given the specialization in natural language processing. In contrast, our approach, using ML models and CEs, offers a more generalizable way to address learner heterogeneity across a wider range of learning activities.

Fifth, we also find that those learners who benefit most from the feedback are those with relatively low or high platform usage. This effect may stem from the feedback's ability to model individual needs across different time periods. For instance, tailored advice might be particularly helpful in cases of low platform usage—often indicative of high procrastination—by guiding learners through minor changes, such as completing test exams shortly before the main exam. Conversely, learners who frequently engage with the platform might benefit from more nuanced guidance throughout their learning process, as they can harvest and implement the feedback as intended. We therefore find that XAI-based feedback supports learners independently of specific SRL-related factors; yet, it seems to particularly help those with relatively low or high platform activity levels.

5.2 Implications for Theory and Practice

Non-personalized behavioral interventions have mainly been shown to be effective for a specific group of learners (Leung et al., 2023; Li et al., 2021; Mousavi et al., 2025). Given that SRL theories posit that learners inherently differ in their use of SRL strategies, it is crucial to take learner heterogeneity into account when providing learner support (Credé and Phillips, 2011; Pintrich, 2000). Otherwise, such support risks exacerbating inequalities by leaving some learners behind.

Our approach addresses this challenge in digital learning environments through adaptive feedback that can equally support learners with varying SRL skills. By showing that the feedback addresses learner heterogeneity, we solidify the theoretical framework of Hattie and Timperley (2007) in the digital space. More specifically, they argue that feedback for students needs to adapt to the individual in order to be effective in learning environments. In this sense, our approach mirrors classroom settings, where instructors adjust their teaching methods to accommodate the varying needs of their learners. Our approach thereby dynamically learns—like instructors do—which learning paths are effective and which are not, adapting its content to the learners' needs.

In order to enable the adaptation to learners, we incorporated domain-specific theory on learner behavior in the feedback development process. In this way, we (a) utilized the insights from SRL theory to understand learner heterogeneity and (b) embedded them

into a digital feedback intervention. More specifically, the development was guided by SRL theory, with the feedback designed to encourage and support SRL strategies (e.g., watching a specific video to reinforce content) considering all phases of the SRL process (e.g., progress monitoring during the performance phase; see Zimmerman and Moylan, 2009). Hence, addressing learner heterogeneity in digital feedback interventions necessitates adopting a sociotechnical view that relies on the insights obtained from established theories on human behavior (see, e.g., Burton-Jones et al., 2021). The implications of established theories such as SRL must not be overlooked in the conceptual development process of feedback interventions. These time-tested theories should therefore be considered more strongly in learner support to promote the transfer of this valuable knowledge into the digital age.

In the search for a suitable approach to implement these insights to address learner heterogeneity, we selected ML and XAI in the form of CEs. Similar to how instructors use their expertise to recommend effective learning actions, ML identifies patterns within learning activities and learners' characteristics, as well as the interplay between both (i.e., activities and characteristics). Building on the patterns obtained from ML, XAI in the form of CEs has the capability to guide learners through context-specific instructions to higher academic success. Prior studies have mainly employed XAI to foster, e.g., trust and fairness in ML model outputs (Haque et al., 2023). Our work goes beyond such applications, as we use XAI to learn from ML to derive targeted insights similar to Senoner et al. (2022) in the context of process quality management. More specifically, we use XAI and CEs to (a) learn from the pattern that ML detects in the data and (b) generate personalized feedback. This also distinguishes our approach from prior IS work in which XAI explanations are surfaced directly to users as part of the feedback they receive (see, e.g., Förster et al., 2025): rather than presenting explanations to learners, we use XAI internally to derive actionable instructions, so that learners act on concrete next steps for their own learning behavior instead of interpreting the explanation themselves.

This approach also makes sense from the perspective of task-technology fit theory that posits that a technology (i.e., here ML and XAI) must fit well with the task (i.e., here learning) it supports to be effective (Goodhue and Thompson, 1995). In other words, supporting learning processes is an overall complex task, which, in turn, requires complex technology to overcome the absence of a teacher in a digital learning environment. Such feedback may be valuable in complex domains that particularly require addressing subject heterogeneity to be most effective. For instance, it is well-known from theories in the environmental and health domain that subjects differ in their response to behavioral interventions (see, e.g., Bryan et al., 2021; Van Valkengoed et al., 2022). Developing interventions based on CEs might therefore maximize the impact of behavioral campaigns.

Beyond the field of higher education, our findings are also relevant to organizational learning. Our studies' context—an IT project management course—closely mirrors an or-

ganizational training (e.g., for preparing and developing project managers), where learners can also differ in their use of SRL strategies (see, e.g., Wan et al., 2012). Specifically, our paper shows that feedback informed by SRL theory and enabled by ML and XAI can tailor learning to individual needs, thereby challenging one-size-fits-all feedback approaches to learning support in organizational digital learning environments. Managers can use these insights to invest in adaptive feedback that strengthens workforce learning. Likewise, software developers can apply our approach to integrate such feedback into firms' digital training platforms and improve outcomes across diverse employee groups.

5.3 Conclusions, Limitations, and Future Research

In this paper, we presented the results of a randomized controlled trial that tested the behavioral responses to feedback generated by ML and XAI. We demonstrated that such an approach can address learner heterogeneity through personalized guidance, thereby increasing online learning activity and improving course performance. These findings highlight the potential of ML and XAI in higher education to create more effective and personalized learning experiences, paving the way for further research across diverse educational contexts.

Despite our best efforts, the study is subject to the following main limitations. To begin, the sample size might be too small to detect further treatment effects, specifically in the analysis on the exam-taking rate or in the moderation analyses. With substantially larger sample sizes, ensuring sufficient statistical power, such effects might become detectable, motivating further research in this area. Moreover, we cannot rule out the possibility that learners in the control and treatment groups discussed the different course design (i.e., absence or presence of the feedback component). Also, participants might have discussed the feedback with each other, with some in the control group potentially realizing that they were less supported within the digital learning environment. Such discussions might have drawn the treatment group's attention more to the feedback, thus increasing its effect, or led to more engaged learners in the control group to make up for the missing information. Apart from that, we cannot credibly argue to what extent the observed effects might extrapolate to courses that differ in content, such as those requiring more mathematical logic and reasoning.

Besides addressing these limitations, we see numerous avenues for future research. In generating feedback, our approach specifically focused on log data that captured certain events from learners (e.g., time spent on specific pages, number of video views). The application of specific SRL strategies, like the generation of mind maps, taking notes, or planning of one's learning, was not part of the online course and therefore was not recorded. Such higher-level strategies are crucial determinants of academic success and would therefore be beneficial to incorporate into XAI-generated feedback in the future.

Relevant psychometric variables could be collected through a survey at the beginning of the course and integrated into the ML model used for predicting course performance. The resulting counterfactuals could then recommend strategies derived from the psychometric variables for individuals to apply. A more general aspect for future IS research is to embrace an idiographic perspective for learner support. This shift reflects a broader trend in various fields (see, e.g., Cram et al., 2024; Hofmann and Hayes, 2019), where traditional non-personalized models are questioned. Instead of a focus on general approaches to mitigate learning issues, such as procrastination, researchers take a process-oriented perspective, acknowledging that there is an excessive number of moderators and mediators involved in the success of a treatment. Here, ML can detect the influences of specific moderators and mediators and, through the application of CEs, researchers might extract novel and more effective interventions. For higher education, CEs could not only be applicable within digital learning environments but could also materialize as a personal digital study tutor that guides learners in selecting courses, managing study schedules, and developing effective study habits. Similarly, a digital study tutor could provide universities with information on where students are not coping well, offering their departments and their staff diagnostic insights into various improvement areas where counterfactuals alone are not sufficient for student support. This might be a very promising future avenue for IS research that focuses on developing smart companions (see, e.g., Schlimbach et al., 2024; Wambsganss et al., 2025).

References

Aucejo, E. M., French, J., Ugalde Araya, M. P., and Zafar, B. (2020). "The impact of COVID-19 on student experiences and expectations: Evidence from a survey." *Journal of Public Economics* 191. DOI: 10.1016/j.jpubeco.2020.104271.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bryan, C. J., Tipton, E., and Yeager, D. S. (2021). "Behavioural science is unlikely to change the world without a heterogeneity revolution." *Nature Human Behaviour* 5 (8), pp. 980–989. DOI: 10.1038/s41562-021-01143-3.

Burton-Jones, A., Butler, B. S., Scott, S. V., and Xu, S. X. (2021). "Next-generation information systems theorizing: A call to action." *MIS Quarterly* 45 (1), pp. 301–314. DOI: 10.25300/MISQ/2021/15434.

Cerezo, R., Esteban, M., Sánchez-Santillán, M., and Núñez, J. C. (2017). "Procrastinating behavior in computer-based learning environments to predict performance: A case study in Moodle." *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.01403.

Coleman, M. D. (2010). "Sunk cost, emotion, and commitment to education." *Current Psychology* 29 (4), pp. 346–356. DOI: 10.1007/s12144-010-9094-6.

Cram, W., D'Arcy, J., and Benlian, A. (2024). "Time will tell: The case for an idiographic approach to behavioral cybersecurity research." *MIS Quarterly* 48 (1), pp. 95–136. DOI: 10.25300/MISQ/2023/17707.

Credé, M. and Phillips, L. A. (2011). "A meta-analytic review of the motivated strategies for learning questionnaire." *Learning and Individual Differences* 21 (4), pp. 337–346. DOI: 10.1016/j.lindif.2011.03.002.

Davis, D., Jivet, I., Kizilcec, R. F., Chen, G., Hauff, C., and Houben, G.-J. (2017). "Follow the successful crowd: Raising MOOC completion rates through social comparison at scale." *Proceedings of the 7th International Learning Analytics & Knowledge Conference*. Vancouver, BC, Canada. DOI: 10.1145/3027385.3027411.

Delen, D., Davazdahemami, B., and Rasouli Dezfouli, E. (2024). "Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework." *Information Systems Frontiers* 26, pp. 641–662. DOI: 10.1007/s10796-023-10397-3.

Everson, H. T. and Tobias, S. (1998). "The ability to estimate knowledge and performance in college: A metacognitive analysis." *Instructional Science* 26 (1–2), pp. 65–79. DOI: 10.1023/a:1003040130125.

Fernández-Loría, C., Provost, F., and Han, X. (2022). "Explaining data-driven decisions made by AI systems: The counterfactual approach." *MIS Quarterly* 46 (3), pp. 1635–1660. DOI: 10.25300/MISQ/2022/16749.

Förster, M., Broder, H. R., Fahr, M. C., Klier, M., and Fink, L. (2025). "Tell me more, tell me more: The impact of explanations on learning from feedback provided by artificial intelligence." *European Journal of Information Systems* 34 (2), pp. 323–345. DOI: 10.1080/0960085X.2024.2404028.

Goodhue, D. L. and Thompson, R. L. (1995). "Task-technology fit and individual performance." *MIS Quarterly* 19 (2), pp. 213–236. DOI: 10.2307/249689.

Gottschalk, F. and Weise, C. (2023). "Digital equity and inclusion in education: An overview of practice and policy in OECD countries." *OECD Education Working Paper No. 299*. Pre-published.

Gray, C. C. and Perkins, D. (2019). "Utilizing early engagement and machine learning to predict student outcomes." *Computers & Education* 131, pp. 22–32. DOI: 10.1016/j.compedu.2018.12.006.

Gupta, S. and Bostrom, R. (2009). "Technology-mediated learning: A comprehensive theoretical model." *Journal of the Association for Information Systems* 10 (9), pp. 686–714. DOI: 10.17705/1jais.00207.

Gupta, S. and Bostrom, R. (2013). "Research note—An investigation of the appropriation of technology-mediated training methods incorporating enactive and collaborative learning." *Information Systems Research* 24 (2), pp. 454–469. DOI: 10.1287/isre.1120.0433.

Haque, A. B., Islam, A. N., and Mikalef, P. (2023). "Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research." *Technological Forecasting and Social Change* 186. DOI: 10.1016/j.techfore.2022.122120.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. 2nd Edition. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84858-7.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge. ISBN: 978-1-134-02412-4.

Hattie, J. and Timperley, H. (2007). "The power of feedback." *Review of Educational Research* 77 (1), pp. 81–112. DOI: 10.3102/003465430298487.

Hoffait, A.-S. and Schyns, M. (2017). "Early detection of university students with potential difficulties." *Decision Support Systems* 101, pp. 1–11. DOI: 10.1016/j.dss.2017.05.003.

Hofmann, S. G. and Hayes, S. C. (2019). "The future of intervention science: Process-based therapy." *Clinical Psychological Science* 7 (1), pp. 37–50. DOI: 10.1177/2167702618772296.

Hope, A. C. A. (1968). "A simplified Monte Carlo significance test procedure." *Journal of the Royal Statistical Society. Series B (Methodological)* 30 (3), pp. 582–598.

Huang, N., Zhang, J., Burtch, G., Li, X., and Chen, P. (2021). "Combating procrastination on massive online open courses via optimal calls to action." *Information Systems Research* 32 (2), pp. 301–317. DOI: 10.1287/isre.2020.0974.

Kalyuga, S. (2007). "Expertise reversal effect and its implications for learner-tailored instruction." *Educational Psychology Review* 19 (4), pp. 509–539. DOI: 10.1007/s10648-007-9054-3.

Kizilcec, R. F., Pérez-Sanagustín, M., and Maldonado, J. J. (2017). "Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses." *Computers & Education* 104, pp. 18–33. DOI: 10.1016/j.compedu.2016.10.001.

Leeds, E., Campbell, S., Baker, H., Ali, R., Brawley, D., and Crisp, J. (2013). "The impact of student retention strategies: An empirical study." *International Journal of Management in Education* 7 (1–2), pp. 22–43. DOI: 10.1504/IJMIE.2013.050812.

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W., and Yue, W. T. (2023). "Could gamification designs enhance online learning through personalization? Lessons from a field experiment." *Information Systems Research* 34 (1), pp. 27–49. DOI: 10.1287/isre.2022.1123.

Li, X. and Zhang, J. (2016). "The effects of monetary incentives and social comparison on MOOC participation: A randomized field experiment." *Proceedings of the International Conference on Information Systems (ICIS 2016)*. Dublin, Ireland.

Li, Z., Wang, G., and Wang, H. J. (2021). "Peer effects in competitive environments: Field experiments on information provision and interventions." *MIS Quarterly* 45 (1), pp. 163–191. DOI: 10.25300/MISQ/2021/16085.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). "From local explanations to global understanding with explainable AI for trees." *Nature Machine Intelligence* 2, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

McCloskey, J. (2011). "Finally, my thesis on academic procrastination." Thesis. Arlington, TX: University of Texas at Arlington. URL: https://mavmatrix.uta.edu/psychology_theses/30/.

Miguéis, V., Freitas, A., Garcia, P. J., and Silva, A. (2018). "Early segmentation of students according to their academic performance: A predictive modelling approach." *Decision Support Systems* 115, pp. 36–51. DOI: 10.1016/j.dss.2018.09.001.

Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267, pp. 1–38. DOI: 10.1016/j.artint.2018.07.007.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations." *Proceedings of the 2020 Conference*

on Fairness, Accountability, and Transparency. Barcelona, Spain. DOI: 10.1145/3351095.3372850.

Mousavi, N., Golara, S., and Bockstedt, J. (2025). "The impact of situational achievement goals on online learning behavior: Results from field experiments." *Information Systems Research* 36 (2), pp. 983–1010. DOI: 10.1287/isre.2022.0353.

Mubarak, A. A., Cao, H., and Zhang, W. (2022). "Prediction of students' early dropout based on their interaction logs in online learning environment." *Interactive Learning Environments* 30 (8), pp. 1414–1433. DOI: 10.1080/10494820.2020.1727529.

Nguyen, A., Tuunanen, T., Gardner, L., and Sheridan, D. (2021). "Design principles for learning analytics information systems in higher education." *European Journal of Information Systems* 30 (5), pp. 541–568. DOI: 10.1080/0960085X.2020.1816144.

Panadero, E. (2017). "A review of self-regulated learning: Six models and four directions for research." *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00422.

Piccoli, G., Rodriguez, J., Palese, B., and Bartosiak, M. L. (2020). "Feedback at scale: Designing for accurate and timely practical digital skills evaluation." *European Journal of Information Systems* 29 (2), pp. 114–133. DOI: 10.1080/0960085X.2019.1701955.

Pintrich, P. R. (2000). "The role of goal orientation in self-regulated learning." In: *Handbook of Self-Regulation*. Elsevier, pp. 451–502. DOI: 10.1016/B978-012109890-2/50043-3.

Pintrich, P. R. (2004). "A conceptual framework for assessing motivation and self-regulated learning in college students." *Educational Psychology Review* 16 (4), pp. 385–407. DOI: 10.1007/s10648-004-0006-x.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1993). "Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ)." *Educational and Psychological Measurement* 53 (3), pp. 801–813. DOI: 10.1177/0013164493053003024.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. Ann Arbor, MI, USA: National Center for Research to Improve Postsecondary Teaching and Learning. URL: https://eric.ed.gov/?id=ED338122.

Pintrich, P. R., Wolters, C. A., and Baxter, G. P. (2000). "Assessing metacognition and self-regulated learning." In: *Issues in the Measurement of Metacognition*. Lincoln: Buros Institute of Mental Measurement, pp. 43–97.

Santhanam, R., Liu, D., and Shen, W.-C. M. (2016). "Research note—gamification of technology-mediated training: Not all competitions are the same." *Information Systems Research* 27 (2), pp. 453–465. DOI: 10.1287/isre.2016.0630.

Santhanam, R., Sasidharan, S., and Webster, J. (2008). "Using self-regulatory learning to enhance e-learning-based information technology training." *Information Systems Research* 19 (1), pp. 26–47. DOI: 10.1287/isre.1070.0141.

Schlimbach, R., Lange, T. C., Wagner, F., Robra-Bissantz, S., and Schoormann, T. (2024). "An educational business model ideation tool – insights from a design science project." *Communications of the Association for Information Systems* 54 (1), pp. 642–661. DOI: 10.17705/1cais.05423.

Schraw, G., Wadkins, T., and Olafson, L. (2007). "Doing the things we do: A grounded theory of academic procrastination." *Journal of Educational Psychology* 99 (1), pp. 12–25. DOI: 10.1037/0022-0663.99.1.12.

Senoner, J., Netland, T., and Feuerriegel, S. (2022). "Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing." *Management Science* 68 (8), pp. 5704–5723. DOI: 10.1287/mnsc.2021.4190.

Steel, P. (2007). "The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure." *Psychological Bulletin* 133 (1), pp. 65–94. DOI: 10.1037/0033-2909.133.1.65.

Steel, P. and Klingsieck, K. B. (2016). "Academic procrastination: Psychological antecedents revisited." *Australian Psychologist* 51 (1), pp. 36–46. DOI: 10.1111/ap.12173.

Štrumbelj, E., Kononenko, I., and Robnik-Šikonja, M. (2009). "Explaining instance classifications with interactions of subsets of feature values." *Data & Knowledge Engineering* 68 (10), pp. 886–904. DOI: 10.1016/j.datak.2009.01.004.

Tuckman, B. W. (2005). "Relations of academic procrastination, rationalizations, and performance in a web course with deadlines." *Psychological Reports* 96 (3), pp. 1015–1021. DOI: 10.2466/pr0.96.3c.1015-1021.

Turel, O. (2025). "To learn or not learn from AI? Unpacking the effects of feedback valence on novel insights recall." *European Journal of Information Systems* 34 (4), pp. 758–779. DOI: 10.1080/0960085X.2024.2426473.

United Nations Department of Economic and Social Affairs (2024). *The sustainable development goals report 2024*. New York, NY, USA: United Nations. URL: https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf.

Van Valkengoed, A. M., Abrahamse, W., and Steg, L. (2022). "To select effective interventions for pro-environmental behaviour change, we need to consider determinants of behaviour." *Nature Human Behaviour* 6 (11), pp. 1482–1492. DOI: 10.1038/s41562-022-01473-w.

Vosniadou, S. (2020). "Bridging secondary and higher education. The importance of self-regulated learning." *European Review* 28 (S1), pp. 94–103. DOI: 10.1017/S1062798720000939.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2), pp. 841–887.

Wambsganss, T., Janson, A., Söllner, M., Koedinger, K., and Leimeister, J. M. (2025). "Improving students' argumentation skills using dynamic machine-learning–based modeling." *Information Systems Research* 36 (1), pp. 474–507. DOI: 10.1287/isre.2021.0615.

Wan, Z., Compeau, D., and Haggerty, N. (2012). "The effects of self-regulated learning processes on e-learning outcomes in organizational settings." *Journal of Management Information Systems* 29 (1), pp. 307–340. DOI: 10.2753/MIS0742-1222290109.

Wolters, C. A. (2003). "Understanding procrastination from a self-regulated learning perspective." *Journal of Educational Psychology* 95 (1), pp. 179–187. DOI: 10.1037/0022-0663.95.1.179.

Yockey, R. (2016). "Validation of the short form of the academic procrastination scale." *Psychological Reports* 118 (1), pp. 171–179. DOI: 10.1177/0033294115626825.

Zhou, M., Lam, K. K. L., and Zhang, Y. (2022). "Metacognition and academic procrastination: A meta-analytical examination." *Journal of Rational-Emotive & Cognitive-Behavior Therapy* 40 (2), pp. 334–368. DOI: 10.1007/s10942-021-00415-1.

Zimmerman, B. J. and Moylan, A. R. (2009). "Self-regulation: Where metacognition and motivation intersect." In: *Handbook of Metacognition in Education*. New York: Routledge, pp. 299–315.

Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course

Supplementary Material

Note: To allow for a blind review process, the version submitted did not yet cite Haag et al. (2023)[1] *(Paper VII of this dissertation), which describes the design and evaluation of the feedback artifact; instead, the relevant parts (section 3–5, adapted and shortened) were attached as supplementary material to that submission. These parts are not reproduced here; see Paper VII for details.*

1 Propensity Score Matching Analyses

For all Propensity Score Matching (PSM) analyses, we used the R packages by Ho et al. (2011) for matching and the package by Arel-Bundock et al. (2024) to compute the average marginal effects for the treatment group. Before matching, we mean-replaced the missing values for seven participants for the variable "procrastination" and for six participants for the variable "use of metacognitive strategies". To estimate the average treatment effect and ensure covariate balance between the treatment and control groups, we employed 1:1 optimal pair matching.

For the initial matching, we created 32 pairs. Due to matching without replacement, two unmatched treatment participants were excluded. Specifically, we matched individuals on the Self-Regulated Learning (SRL) variables that reflect procrastination behavior and the use of metacognitive strategies, as well as the number of weeks participants were online during the intervention:

$$T_i^{\text{Feedback}} \sim SRL_i^{\text{Proc.}} + SRL_i^{\text{Metacog.}} + W_i^{\text{Online}}. \tag{1}$$

[1] F. Haag et al. (2023). "Addressing Learners' Heterogeneity in Higher Education: An Explainable AI-based Feedback Artifact for Digital Learning Environments." *Proceedings of the 18th International Conference on Wirtschaftsinformatik*. Paderborn, Germany

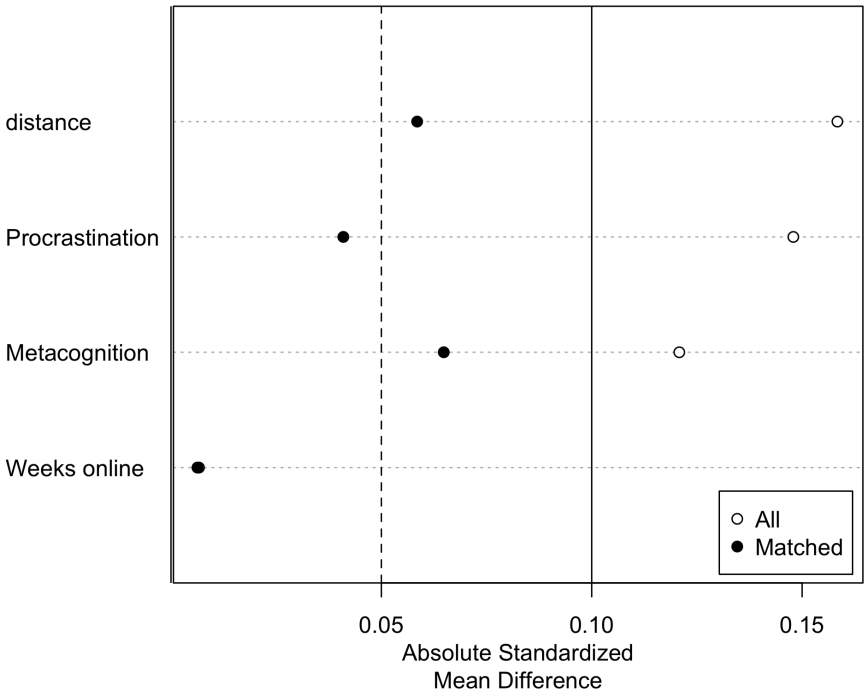


Figure 1: Absolute standardized mean differences before (All) and after matching (Matched) for the initial analysis.

After matching, all covariates display an absolute standardized mean difference below 0.1 (Figure 1). The *distance* variable represents the propensity score based on the covariates SRL_i and W_i^{Online}. The propensity score captures the probability of treatment assignment given the covariates. In this analysis, the average absolute standardized mean difference of the distance variable across treatment and control groups was below 0.1 after matching, indicating an excellent level of covariate balance (Rubin, 2001). This suggests that the treated and control groups are sufficiently similar with respect to the covariates used for matching. In addition, a Kolmogorov-Smirnov test indicates that the propensity score distributions for the treatment and control group are similar ($p = 0.160$).

Finally, we weighted the outcome regression model (see Equation 6 in the paper) for effect estimation by the propensity score weights obtained during matching. Our analyses thereby employ clustered standard errors to adjust for the matched pair structure.

1.1 Moderation analysis: Procrastination

To investigate the moderating effect of procrastination on the treatment outcome, we conducted a median split on the procrastination variable, categorizing participants into *high* and *low* procrastination groups. Deviating from the analysis above, we performed exact matching on the procrastination group (i.e., high and low) variable to ensure that participants within pairs exhibited the same procrastination category. The matching model also included the number of weeks participants were online during the intervention and the

use of metacognitive strategies as additional covariates. In addition, we included the interaction term between the procrastination group and the weeks participants were online because there might be a plausible relationship between these two variables. Specifically, participants with higher procrastination tendencies may exhibit different patterns of platform usage over time, such as delayed or inconsistent engagement. Accounting for this interaction allows us to better capture the joint influence of procrastination and online activity on the treatment assignment and subsequent outcomes:

$$T_i^{\text{Feedback}} \sim SRL_i^{\text{Proc. (median split)}} + W_i^{\text{Online}} + SRL_i^{\text{Proc. (median split)}} \times W_i^{\text{Online}} + SRL_i^{\text{Metacog.}}. \quad (2)$$

The matching resulted again in 32 pairs and an excellent covariate balance (Rubin, 2001), with all absolute standardized mean differences below 0.1 after matching (Figure 2).

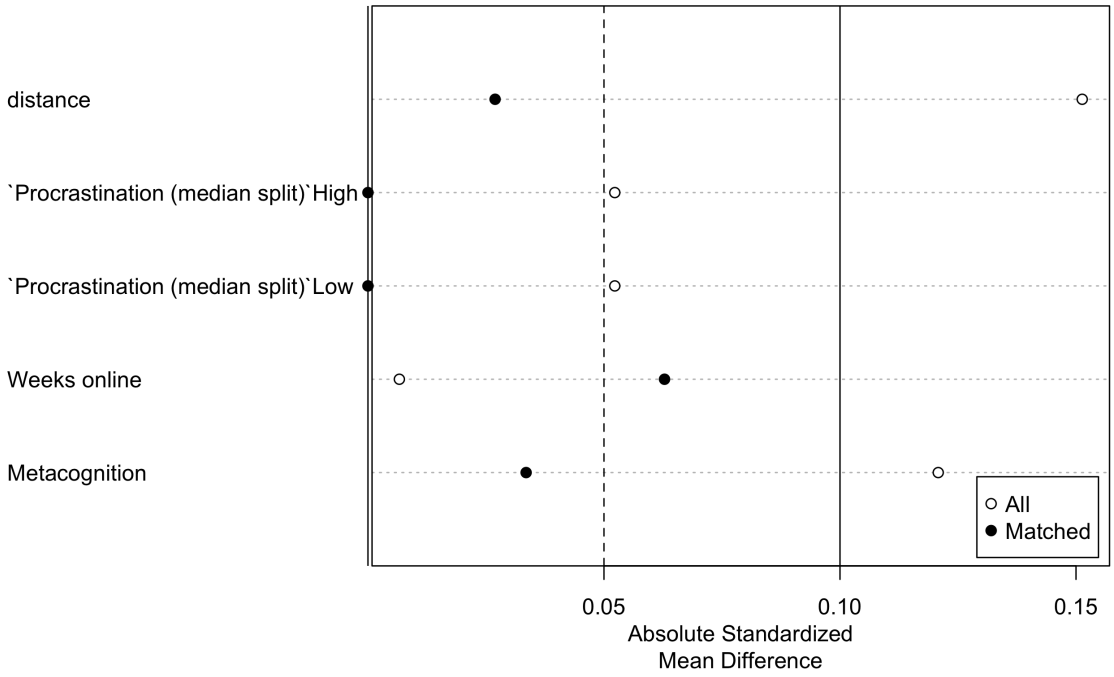


Figure 2: Absolute standardized mean differences before (All) and after matching (Matched) for the moderation analysis on procrastination.

To examine the conditional effects, we computed average treatment effects separately for participants with *high* and *low* procrastination levels and also performed moderation analysis by examining group differences. For this purpose, we fit a weighted linear regression model including an interaction term between the treatment group and procrastination group (i.e., moderator):

$$\begin{aligned} y_i = \alpha + \beta_0 T_i^{\text{Feedback}} + \beta_1 SRL_i^{\text{Proc. (median split)}} \\ + \beta_2 T_i^{\text{Feedback}} \times SRL_i^{\text{Proc. (median split)}} + \varepsilon_i. \end{aligned} \quad (3)$$

1.2 Moderation analysis: Metacognition

For the moderation analysis on metacognition, we applied a procedure similar to that used for the procrastination analysis. A median split categorized participants into the levels *high* and *low* use of metacognitive strategies. Again, our analysis relied on exact matching on the metacognition group variable, while procrastination and the number of weeks online served as additional covariates:

$$T_i^{\text{Feedback}} \sim SRL_i^{\text{Metacog. (median split)}} + W_i^{\text{Online}} + SRL_i^{\text{Proc.}}. \tag{4}$$

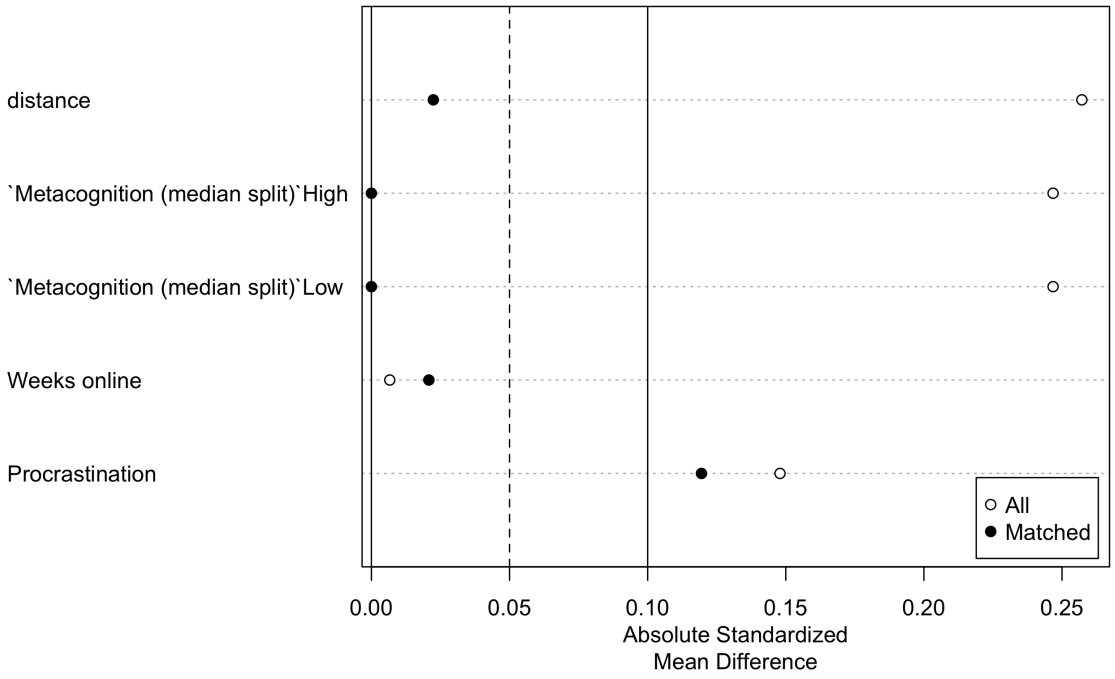


Figure 3: Absolute standardized mean differences before (All) and after matching (Matched) for the moderation analysis on metacognition.

The matching results in 29 pairs (three participants from the control group and five participants from the treatment group were excluded) and an appropriate covariate balance, with all absolute standardized mean differences below 0.1 for all variables except for procrastination (Figure 3). The matching results can still be considered appropriate, as we can see an improvement compared to the pre-matching imbalance. Additionally, all other covariates achieve excellent balance, ensuring that any residual imbalance in procrastination is unlikely to bias the results.

Similarly to the procrastination analysis, we calculated the effects separately for the levels *high* and *low* and analyzed subgroup differences using the following outcome model:

$$\begin{aligned} y_i = \alpha + \beta_0 T_i^{\text{Feedback}} + \beta_1 SRL_i^{\text{Metacog. (median split)}} \\ + \beta_2 T_i^{\text{Feedback}} \times SRL_i^{\text{Metacog. (median split)}} + \varepsilon_i. \end{aligned} \tag{5}$$

References

Arel-Bundock, V., Greifer, N., and Heiss, A. (2024). "How to interpret statistical models using marginaleffects for R and Python." *Journal of Statistical Software* 111 (9), pp. 1–32. DOI: 10.18637/jss.v111.i09.

Ho, D., Imai, K., King, G., and Stuart, E. A. (2011). "MatchIt: Nonparametric preprocessing for parametric causal inference." *Journal of Statistical Software* 42 (8), pp. 1–28. DOI: 10.18637/jss.v042.i08.

Rubin, D. B. (2001). "Using propensity scores to help design observational studies: Application to the tobacco litigation." *Health Services and Outcomes Research Methodology* 2, pp. 169–188. DOI: 10.1023/a:1020363010465.

Appendix

Publications

Articles (Peer reviewed)

Haag, F. (2026). "How Explanations from XAI-based Decision Support Affect Human Task Performance: A Meta-Analysis." *Journal of Decision Systems* 35 (1). DOI: 10.1080/12460125.2026.2616693.

Haag, F., Hopf, K., and Staake, T. (2026). "Validating Explainer Methods: A Functionally Grounded Approach for Numerical Forecasting." *Journal of Forecasting* 45 (2), pp. 819–836. DOI: 10.1002/for.70060.

Haag, F. (2025). "The Effect of Explainable AI-based Decision Support on Human Task Performance: A Meta-Analysis." *Proceedings of the 23rd Annual Pre-ICIS Workshop on HCI Research in MIS*. Bangkok, Thailand.

May, M., Hopf, K., Haag, F., Staake, T., and Wortmann, F. (2025). "Fostering active student engagement in flipped classroom teaching with social normative feedback." *Proceedings of the 20th International Conference on Wirtschaftsinformatik*. Muenster, Germany. Forthcoming.

Bayer, D. R., Haag, F., Pruckner, M., and Hopf, K. (2024). "Electricity Demand Forecasting in Future Grid States: A Digital Twin-Based Simulation Study." *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*. Bol and Split, Croatia. DOI: 10.23919/splitech61897.2024.10612563.

Giacomazzi, E., Haag, F., and Hopf, K. (2023). "Short-Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources." *Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy)*. Orlando, FL, USA. DOI: 10.1145/3575813.3597345.

Günther, S. A., Haag, F., Hopf, K., Handschuh, P., Klose, M., and Staake, T. (2023). "A Feedback Component That Leverages Counterfactual Explanations for Smart Learning Support." In: *Digitale Kulturen der Lehre entwickeln – Rahmenbedingungen, Konzepte und Werkzeuge*. Edited by L. Mrohs, J. Franz, D. Herrmann, K. Lindner, and T. Staake. Perspektiven der Hochschuldidaktik. Wiesbaden, Germany: Springer VS.

Haag, F., Günther, S. A., Hopf, K., Handschuh, P., Klose, M., and Staake, T. (2023). "Addressing Learners' Heterogeneity in Higher Education: An Explainable AI-based Feedback Artifact for Digital Learning Environments." *Proceedings of the 18th International Conference on Wirtschaftsinformatik*. Paderborn, Germany.

Haag, F., Stingl, C., Zerfass, K., Hopf, K., and Staake, T. (2023). "Overcoming Anchoring Bias: The Potential of AI and XAI-based Decision Support." *Proceedings of the 44th International Conference on Information Systems*. Hyderabad, India.

Haag, F., Hopf, K., Menelau Vasconcelos, P., and Staake, T. (2022). "Augmented Cross-Selling Through Explainable AI—A Case From Energy Retailing." *Proceedings of the 30th European Conference on Information Systems*. Timișoara, Romania.

Wastensteiner, J., Weiss, T. M., Haag, F., and Hopf, K. (2021). "Explainable AI for Tailored Electricity Consumption Feedback – An Experimental Evaluation of Visualizations." *Proceedings of the 29th European Conference on Information Systems*. Marrakech, Morocco (virtual).

Manuscripts

Haag, F., Günther, S. A., Handschuh, P., Klose, M., Hopf, K., and Staake, T. (submitted). "Counterfactual Explanation-Based Feedback for Personalized Learning Support: Evidence from an IT Project Management Course." Submitted to the European Journal of Information Systems.