Machine learning: Essays on governance and value-creation

Doctoral Dissertation

Submitted to the

Faculty of Business and Economics at TU Dortmund University

in partial fulfillment of the requirements for the degree of

Doctor rerum politicarum (Dr. rer. pol.)

By

Julian Sengewald

Dortmund, Februar 2026

Table of contents

A: Introduction to the Thesis

"I beseech you, in the bowels of Christ, think it possible you may be mistaken."

Oliver Cromwell

Introduction

A.1 Motivation

Organizations are increasingly turning to machine learning (ML) systems in order to automate tasks that were carried out by humans or to provide new services. This is occurring for a number of reasons. On the one hand, access to data has become cheaper and easier to collect. Since plenty of data is necessary in order to develop ML systems, this development has allowed organizations to leverage the benefits that ML systems have to offer (Abbasi et al., 2016).

In the first two decades of the new ML revolution, there was primarily a scientific and industrial interest in developing technologies that were capable of handling large volumes of data (big data analytics) as well as in further developing technical aspects of machine learning technology (deep learning). However, given the increased adoption of ML systems, academics and industry are now turning more of their research efforts towards investigating the impact that ML systems have. Ethical questions such as privacy awareness and non-discrimination of the affected individuals are being studied, and legislative bodies around the world are pressured to develop new regulations. Similarly, organizations, such as enterprises, are pivoting their AI/ML strategies toward business operations, shifting from predictive to prescriptive analytics. This evolution stems from maturing AI/ML technologies and the growing demand for data-driven decision-making. Where before there was mostly work on predictive analytics, enterprises and academia are now increasingly turning efforts towards understanding how the use of ML and AI can help them to make better decisions and what the implications of AI/ML use are.

The use of AI and ML in business organizations is informed by cross-disciplinary discourse. While computer science and engineering predominantly drive AI/ML advancement with emphasis on technical dimensions, these developments primarily maintain an algorithmic focus. AI/ML from an organizational perspective must address the broader societal, organizational, and individual implications of these technologies.

AI/ML technology must support business functions to align with the organization's overall purpose-driven objectives. The purpose-driven objectives (e.g., profitability) ensure that the organization remains sustainable as AI/ML technology contributes positively to the financial bottom line. Historically, managers achieve the profitability objectives by structuring the organization into business functions and organizing the work that has to be performed within the business function using resources such as human workers, knowledge, and technology. The unique characteristic of AI/ML lies in its ability to automate activities that are challenging for

humans to explicitly describe and excel at tasks where humans are disadvantaged due to AI/ML's speed and memory capacity. AI/ML has the potential to automate or augment human work in areas previously difficult to formalize using traditional algorithms. AI/ML is typically developed by computer scientists who lack a stake in the business functions and operated by specialized staff with a deep understanding of their daily tasks but limited knowledge of AI/ML capabilities (Samtani et al., 2023).

Organizations operate within ethical boundaries, both explicit (juridical) and implicit (corporate responsibility). The purpose-driven objective implies that organizational AI/ML deployments fulfill specific objectives such as increased operational efficiency or improved decision-making. Such specific objectives constitute the primary evaluation criteria for AI/ML systems within organizational contexts. Explicit regulations describe legal AI/ML applications in organizations. Implicit regulations, though not codified in law, represent socially legitimate applications of AI/ML technologies in response to ongoing discourse regarding their potential risks. Organizations do not employ AI/ML in isolation; they integrate these technologies into their products and services, creating outcomes that directly affect customers and employees. As a result, organizational AI/ML implementation necessitates ethical assessment to address concerns, including privacy violations, algorithmic bias, and social implications of these technologies. These ethical assessments of AI/ML implementations form the foundation of responsible AI governance within organizations.

A.2 Aims

The thesis examines ML and AI in organizational contexts, conceptualizing them as technologies that enable information systems within organizations. For this research, organizations are defined as multi-entity structures operating within a defined scope, where employees perform work that produces services for internal or external consumption (Alter, 2008). In this thesis, the use of AI/ML in organizations that use AI/ML is the primary focus. In particular, AI/ML technologies are conceptualized as being a technological component forming part of the 'work system' definition (Alter, 2008). This framework includes both the technology and the broader organizational context. This perspective raises important considerations such as AI governance, regulatory requirements, and instrumental orientations that emphasize cost-benefit analysis of AI/ML implementations in organizational settings.

These organizational AI/ML themes inform the overarching aim of this dissertation:

"Aid organizations in using AI/ML that is informed by technology and organizational aspects."

Although algorithmic decision-making (ADM) systems, where the algorithms are based on ML and AI as understood in this thesis, are abstract constructs, they may materialize concrete impacts (including harms) in society. One such harm is unfairness (i.e., discrimination) in the decisions derived from the ADM. Fairness in ADM-ML systems refers to how predictions affect decisions involving particular individuals. When these decisions have significant outcomes and benefits, e.g., credit approval, university admission, and resume pre-screening, the unjust allocation of these outcomes/benefits is problematic. For instance, systems that unfairly advantage one subgroup over another are characterized as exhibiting algorithmic bias. Even in partially automated contexts, ML systems maintain discretion over final decisions, which increases in fully automated implementations. Therefore, all ML system deployments require monitoring for bias and fairness concerns, forming part of ML system auditing within AI/ML governance. In addition to bias and fairness, responsible ML must also address privacy considerations (Papagiannidis et al., 2025). ML systems require data as input, and, as such, if that data is personal data, they can create privacy harms. At the same time, privacy management and mitigating the risks of privacy harms are also challenges. Privacy management often employs "masking", the deliberate alteration of data to enhance privacy. This process modifies individual records, so they no longer match the original values exactly. However, the question is how these alterations affect the performance, utility, and fairness. ADM-ML systems can also have undesirable consequences, including the automation of human decisions and the potential loss of skills when decisions are increasingly made by machines. The de-skilling of human expertise has severe consequences, as humans may be less able to monitor and maintain ML systems. The question is therefore how to effectively combine human and machine intelligence without relying too heavily on either.

Fairness, privacy, and skill-erosion from automation are all relevant questions involving AI/ML governance from a managerial perspective. My first research question (RQ) is therefore:

RQ1: *How can ADM-ML be ensured to act responsibly on humans?*

Given the broad nature of RO1 as a research objective, involving human and computerized-technical components, I have divided the meta-objective RO1 into two sub-objectives depending on the primary unit of study. The first sub-objective RO1a concerns empirical work involving primarily humans as the unit of study. The second sub-objective RO2b involves primarily the computerized-technical component as a unit of study. Sub-objectives RO1a and RO1b are closely linked through the meta-objective arising from RO1:

RQ1a: *How can responsible ADM-ML be characterized empirically?*

and

RQ1b: *What is the behavior of complex ML systems in a sociotechnical dimension?*

The motivation behind RO1 is obtaining normative knowledge about how components of AI/ML systems interrelate with respect to ethical outcomes. This foundational understanding directly informs both RO1a and RO1b, which, through distinct research approaches, investigate how these systems can be designed to align with human values and be deployed responsibly. A critical aspect of this research involves identifying and mitigating unintended consequences and negative impacts of AI/ML systems, particularly addressing issues such as bias, discrimination, and privacy concerns. These considerations are essential for ensuring that AI/ML systems are socially sustainable, as they directly affect an organization's reputation and regulatory compliance (Enholm et al., 2022). Furthermore, this compliance perspective is inherently forward-looking, since human-facing AI/ML systems undergo continuous social evaluation after deployment, ultimately shaping both societal norms and regulatory frameworks (Papagiannidis et al., 2025). Consequently, responsible AI/ML usage fundamentally serves organizational interests by preventing potential backlash. In this context, AI governance, to which RO1 significantly contributes, functions as the organizational mechanism that ensures deployed AI/ML systems operate sustainably within societal boundaries. By systematically embedding ethical considerations into AI development processes, organizations can maintain public trust while preserving the operational flexibility that might otherwise be constrained by reactive regulation prompted by societal concerns.

The motivation behind RQ2 is obtaining actionable knowledge about how organizations can systematically derive business value from AI/ML technologies beyond computer science. A key aspect involves translating complex business problems into analytical frameworks that can be addressed by AI/ML systems, particularly by focusing on prescriptive analytics and decision optimization methodologies (Arnott & Pervan, 2005, p. 68). This value-creation perspective is inherently integrative, as ADM-AI/ML systems bridge technical capabilities with business objectives, reshaping organizational decision-making processes and performance metrics. My last research question therefore addresses the objective:

RQ2: *How can organizations derive value from ML/AI systems from aligning them with their business objectives?*

Business-oriented AI/ML development, to which RQ2 contributes, focuses on the value-creation or derivation of the developed AI/ML systems. Value in this context ranges from improved organizational decision-making to financial returns that drive profit. The pursuit of

this research goal involves identifying business problems, abstracting them into analytical solutions addressable by AI/ML and data, and implementing and evaluating AI/ML systems. The analytical solutions developed in response to RO2 serve to improve organizational value derivation from ML/AI systems, as demonstrated in parts 3 and 4 of this thesis.

A.3 Structure of the Thesis

The thesis is divided into three parts: A, B and C. Part A provides an introduction to the thesis, Part B presents each contribution, and Part C concludes the thesis.

Part B contains *nine research contributions* and is organized into *three thematic collections*, each addressing distinct aspects of AI and ML applications. The first *Collection I.* "Ethical Foundations" comprises three chapters focused on responsible AI/ML. It provides a descriptive and constructive problematization of fairness in AI/ML through empirical survey research. The *Collection II.* "Understanding Inequity in ML Systems: A Data-Centric Approach" consists of two chapters. These two chapters examine data dilemmas that cause unfairness in ML systems. While Part I. establishes theoretical foundations, Part II. takes a technical approach to exploring sources of unfairness, aligning with the ethical perspective established earlier. The final *Collection III.* "Optimization and Management in AI Systems" contains three chapters discussing how organizations can apply AI and ML systems to optimize their operations. Unlike the general perspectives in Parts I. and II., these chapters propose specific ML artifacts for organizational process optimization. The focus shifts from the customer-facing, responsible ML design (established in Collection I. and II.) to operation-oriented optimization for profitability.

The structuring in the parts in this cumulative dissertation was decided on the basis of thematic and application alignment of the chapters in each part. The progression from ethical foundations to practical applications reflects a holistic approach to understanding the challenges and opportunities presented by AI/ML for ADM in various domains.

With respect to the formulated research objectives, Collection I. revolves around RO1a whereas Collection II. revolves around RO1b. In sum, RO1 consists of four papers, which are arranged over two thematic collections. RO2 comprises four papers, which are organized in collection III.

The collections are not strictly sequential and can be read independently. The thematic organization serves to highlight the interconnectedness of the research contributions while allowing for flexibility in reading the individual chapters. Each collection and chapter is self-contained and can be understood without prior knowledge of the other collection or chapters. However, the chapters within each part are presented in the order of their subsequent

contribution. The first chapter of each part serves as an introduction to a broader topic, while subsequent chapters delve into specific aspects. Despite the flexibility in the arrangement of the parts and chapters, I encourage the reader to read collection I before part II.

Each chapter in this cumulative dissertation is presented as a standalone paper, with its own introduction, literature review, methodology, results, and discussion sections. The publication outlets for each chapter are shown in Table A.1, which provides a summary of the research contributions and their significance in the context of the thesis. All chapters are published in peer-reviewed conference proceedings (except chapter 5, which is a working paper), ensuring the quality and rigor of the research presented. The outlets belong to the domains of information systems and the information systems section at computer science conferences (i.e. HICCS). In Table A.1, each contribution is assigned a classification according to the VHB4 ranking[1] and WKWI guidance list[2]. The chapters are written in a consistent style and format. The references are provided at the end of the dissertation to lessen the extent of redundant bibliography entries.

A more detailed overview of the chapters and their contributions is provided in the following sections. In that section A.6 I also discuss how each chapter contributes to the overall research objectives of the thesis and how these contributions are related to information systems research (ISR).

[1]https://www.vhbonline.org/fileadmin/vhb/Services/vhb-rating/WI/VHB_Rating_2024_Area_rating_WI.pdf

[2] https://link.springer.com/article/10.1365/s11576-008-0040-2

Table A.1: Overview of contributions and their publication outlets

C	RQ	Citation	Outlet	VHB4	WKWI
1	1	Sengewald, J., & Lackes, R. (2022). Fair Engineering of Machine Learning Systems–Lessons Learned from a Literature Review. Proceedings of the 55th Hawaii International Conference on System Sciences	HICCS	B	B
2	1	Sengewald, Julian; Schlichter, Anissa; Siepermann, Markus; and Lackes, Richard, "Human perceptions of fairness: a survey experiment" (2023). Wirtschaftsinformatik 2023 Proceedings. 72.♦	WI	B	A
3	1	Sengewald, Julian; Wilz, Mathis; and Lackes, Richard, "AI-Assisted Learning Feedback: Should Gen-AI Feedback Be Restricted to Improve Learning Success? A Pilot Study in a SQL Lecture" (2024). ECIS 2024 Proceedings. 12.♠	ECIS	A	A
4	2	Sengewald, J., & Lackes, R. (2021). The impact of the 'right to be forgotten' on algorithmic fairness. In Perspectives in Business Informatics Research: 20th International Conference on Business Informatics Research, BIR 2021, Vienna, Austria, September 22–24, 2021, Proceedings 20 (pp. 204-218). Springer International Publishing.	LNBIP	(C*)	nr
5	2	Sengewald, J., & Lackes, R. (2025). Balancing Privacy, Fairness, and Utility in Data Sharing: Synthetic vs. Perturbative Approaches a Robust Assessment.	*Working Paper*.		
6	3	Sengewald, Julian and Lackes, Richard, "Prescriptive Analytics in Procurement: Reducing Process Costs" (2022). Wirtschaftsinformatik 2022 Proceedings. 5.	WI	B	A
7	3	Sengewald, J., & Lackes, R. (2024, May). A Multi-objective Particle Swarm Optimization Framework for Operations Management. In Wuhan International Conference on E-business (pp. 444-455). Cham: Springer Nature Switzerland.	LNBIP	(C*)	nr
8	3	Sengewald, Julian and Lackes, Richard, "Robo-Advisory and Algorithmic Trading via Evolutionary	AMCIS	C	B

		Discretization and Rule-Mining" (2024). AMCIS 2024 Proceedings. 12.±			
9	1	Sengewald, Julian and Lackes, Richard, "Bike-Sharing Station Placement: Spatial Analysis and Data Mining of Network Design Characteristics" (2024). AMCIS 2024 Proceedings. 10.	AMCIS	C	B

'*': While these conferences are not explicitly listed in the VHB conference proceedings ranking, other conference proceedings published in Lecture Notes in Business Information Processing (LNBIP) are assigned a 'C' ranking and LNBIP was also recognized as 'C' in the earlier Jourqual 3 ranking system.

'±': Best paper nominee.

'♦': For this contribution Anissa Schlichter provided the data collection and also co-developed the experimental design. Markus Siepermann provided helpful feedback on writing structure and interpretation of results.

'♠': For this contribution Mathis Wilz provided input for the data collection (mock exam and pre-testing it), assisted in proofreading and made helpful suggestions.

A.4 Conceptualization of Research Artifact

Narrowing down the context of information systems (IS) I refer to the definition of "work system" (Alter, 2008). A "work system" is characterized as a system comprising humans and machines within an organizational context (Alter, 2008). Humans and machines perform information-based work, delivering services to internal or external customers (Alter, 2008).

Using the work system definition and its organizational context implies questions such as how organizations can create value from data. It also raises questions about how organizations must implement their customer-facing ADM applications to generate business value, which includes direct and indirect economic outcomes and IT adoption (Ghasemaghaei & Kordzadeh, 2025; Kordzadeh & Ghasemaghaei, 2021).

The research agenda for these IS revolves around the sociotechnical model in information systems research (ISR). The sociotechnical model, as conceptualized in ISR, views information systems as comprising two interrelated components: the social and the technical, which function as mutually interacting parts of an integrated system (Sarker et al., 2019). In this thesis, the sociotechnical systems studied are those that augment or complement human work. One such system is algorithmic decision-making (ADM). ADM refers to the various forms of interaction between algorithms and human decisions, where human decisions may be augmented, automated, or combined in a hybrid form (Ågerfalk et al., 2022, p. 425; Rai et al., 2019, p. 5). The scope of ADM in this thesis ranges from applications where algorithms have high agency over human decision-making, such as prescriptive analytics that provide direct decisions, to low-agency cases where algorithms inform humans and provide perspective, such as large

language models and XAI, while still influencing the decisions ultimately made by humans to some extent (Bauer et al., 2023). The algorithm in ADM refers in this thesis to those algorithms that are based on ML and artificial intelligence (AI).

ML constitutes the technical component of these ADM systems, broadly understood as algorithms that learn principles and patterns for decision-making from data (Abbasi et al., 2016; Samtani et al., 2023). While many IS publications have studied ADM and technology, ML introduces unique features such as autonomy, adaptive learning, and non-scrutiny to the inner workings (Berente et al., 2021). These features differentiate ML/AI from conventional technology, justifying its study as a distinct category. The social component includes individuals and organizations that use or are affected by ADM systems. This aspect includes individual customers and employee groups within organizations that utilize information systems. Additionally, the social component is embedded in the work systems definition of IS (Alter, 2008). The social component is addressed by examining how the technical affects the user of IS (Ghasemaghaei & Kordzadeh, 2025).

A.5 Research Methods Applied

This thesis employs both behavioral IS research (Maruping et al., 2025) and design-science-oriented ML research (Arnott & Pervan, 2016; Padmanabhan et al., 2022), thereby addressing the sociotechnical nature of IS through multiple complementary approaches (Sarker et al., 2019). While RO1 is mainly addressed using behavioral research methods and partly through computer experiments (M. Zhang & Gable, 2014), RO2 is primarily addressed using design-science approaches.

Behavioral research approaches employ quantitative methods such as surveys and experiments (Maruping et al., 2025). Behavioral research aims to understand phenomena by measuring and quantifying them, particularly human behavior in relation to technology in the IS context (Maruping et al., 2025). For instance, contribution C2 uses a survey experiment in the form of a vignette study with a within-subjects design to study social norms regarding fairness perceptions of ML decisions. Similarly, contribution C3 employs a randomized-control study (between-subject design) in the form of a user study, a method also typical in research conducted at large tech companies.

For literature reviews, several approaches exist. A systematic literature review (SLR) involves a structured search process documenting which literature was selected. A SLR is appropriate when the goal is a comprehensive summary of existing literature (Rowe, 2014). SLR also synthesizes prior knowledge, requiring sufficient previous research in the field (Rowe, 2014).

I selected SLR for C1 to understand current knowledge on fairness perceptions in AI/ML contexts. Contribution C2 expands on this literature review with an experimental study.

Computational experimentation represents another research method employed. In C4 and C5, I used computational experimentation to describe and test (M. Zhang & Gable, 2014) interactions between social and technical environments, focusing on process parameters such as user behavior and technical implementations for data management. Specifically, these contributions use simulation methods to study ML in complex systems (Padmanabhan et al., 2022).

This thesis also employs artifact development within decision support systems (Arnott & Pervan, 2016). Such development is shaped by the application environment – comprising organizational contexts, technological infrastructure, and stakeholders – for which the method is designed (Samtani et al., 2023). These artifacts often address operational aspects of business problems through concrete instantiations of information systems (Arnott & Pervan, 2016). Custom-developed ML methods are a specific form of this design approach (Padmanabhan et al., 2022). From a technical perspective, ML development in IS encompasses specialized data encoding and learning strategies addressing domain-specific challenges (Padmanabhan et al., 2022). ML method development requires an understanding of the problem domain to develop ML methods tailored to specific operational processes in organizations (Padmanabhan et al., 2022). These methods are also evaluated using appropriate domain-specific performance metrics (Padmanabhan et al., 2022). Importantly, ML method development in ISR aims not merely at predictive results – as often seen in computer science – but primarily at positive impacts on business and societal outcomes (Samtani et al., 2023), enhancing managerial decision-making and addressing real-world organizational needs (Arnott & Pervan, 2016).

This thesis uses the artifact development method for ML in business to create several different artifacts, all deliberately designed to impact business and societal outcomes. For example, C6 draws on organizational behavior theory of procurement agents and the specific structure of the procurement sourcing process to develop a custom ML method that guides sourcing decisions. C6 suggests a custom deep learning procedure (Samtani et al., 2023) to directly optimize organizational outcomes in operations management. Thus, contributions C6 and C7 are deeply informed by organizational processes and their configurations. Contributions C8 and C9 offer domain-specific contributions (Padmanabhan et al., 2022). All contributions in C6-9 are centered around value-creation, whether in terms of actual profit or improved decision-making, which ultimately leads to increased profit. Table A.2 summarizes the main research methods used for each contribution in this thesis.

Table A.2: Overview of research methods used in the contributions

Contribution (C)	1	2	3	4	5	6	7	8	9	Reference
Literature (e.g., SLR)	x	x	x	x	x	x	x	x	x	(M. Zhang & Gable, 2014)
Behavioral		x	x							(Maruping et al., 2025)
ML in complex systems				x	x					(Padmanabhan et al., 2022)
Artifact development						x	x	x	x	(Arnott & Pervan, 2016; Samtani et al., 2023)

A.6 Overview of Contributions, Relations among Contributions and Embedding in ISR

This thesis advances the sociotechnical knowledge base in ISR, focusing specifically on how organizations can deploy ADM within the IS-'work systems' definition. Three main themes emerge regarding how these ADMs: 1. affect individuals and society (Contributions 1-3, 5), and are also shaped by them (Contribution 4), 2. operate under societal regulations and norms while considering the well-being of individuals affected by these systems (Contributions 2-5), 3. can support human well-being and provide value to organizations (Contributions 6-9). These objectives align with the sociotechnical perspective's joint-optimization paradigm (Sarker et al., 2019).

IS publications can be positioned within the sociotechnical model on a continuum between technical and social facets, either at their intersection or leaning towards one side being more socio-/technic-centric (Sarker et al., 2019, p. 707). This research aligns with this continuum model by presenting nine contributions positioned across the technical-social spectrum. The methodological approaches in IS research can be classified as following either a behavioral or design-science orientation, where the behavioral orientation is used to inform the design of new artifacts, and the design of new artifacts seeks to augment human knowledge by creating new testable theories (Niederman & March, 2012). This thesis employs behavioral IS research (Maruping et al., 2025) and design-oriented ML research (Arnott & Pervan, 2016; Padmanabhan et al., 2022), thereby addressing the sociotechnical nature of IS through multiple complementary approaches.

In a first set of contributions (C1-3) a predominantly social lens on AI/ML is applied. Contributions 1 and 2 of this thesis explore the question of how to design ADM decision-making processes to be perceived as fair. Given the progress of ML and AI, more and more decision-making processes are completely automated or partially supported by machine decision-making. Contribution 1 reviews the literature on fairness perceptions in ADM. It finds a lack of consensus on suitable fairness notions and of relevant application studies in IS organizational contexts. Contribution 1 also critically engages with the accumulated knowledge from previous empirical literature on algorithmic fairness. It specifically problematizes experimental procedures and contextual factors used in prior research, examining them in light of and with aid of the obtained results. This critical analysis helps to better understand the emerging phenomenon of ADM fairness, as technical solutions alone cannot address fairness without considering the broader social context in which algorithms operate and are perceived. Contribution 2 leverages an experimental survey to study the design of fair ML for ADM-based recruiting. Contributions 1 and 2 thus inform on how to develop ADM-based systems based on ML adhering to ethical principles. In particular, the focus is on which *fairness* metrics are most appropriate in ADM-based recruiting. Algorithmic outcomes refer to the decisions that these systems make and the allocation of benefits that are attached to these decisions. In the end, potential unfairness in such ADM-based systems arises through the consequences that the affected individuals face, as these consequences may have been different if the ADM-based system had made a different decision. This notion of fairness belongs to improving distributive fairness of such systems and is oriented on the output of these systems. The orientation on the output of such systems can be used to address a wide range of sources of bias in ML in a post-hoc manner (Mehrabi et al., 2022). Contribution 7 belongs to a broader class of ML-enabled ADMs as research is just beginning to understand how ML can harm humankind, and Contribution 3 is exploring a mechanism by which such harms may realize. While Contributions 1 and 2 focus on (un-)fairness as an aspect of responsible AI, Contribution 3 covers AI having a potential deskilling effect on human expertise. Such a deskilling effect has the potential to provide harm e.g. to knowledge workers and students. In particular, Contribution 3 examines the implications of delegating cognitive tasks to LLMs and the potential delayed impact on human performance. Contribution 3 as well explores a user interface design to mitigate the negative consequences of cognitive delegation. Contributions 1, 2, and 3 thus provide a dark-side perspective on AI (Ågerfalk et al., 2022). These contributions apply a predominantly social lens on technology (Sarker et al., 2019). Contributions 1, 2 and 7 form a body of research that aligns with the humanistic goals of the

sociotechnical model (Sarker et al., 2019), exploring how technology may harm humanity and how to design responsible technology.

ML is built upon a foundation of data. The increasing volume of data facilitates the development of ML applications and contributes to the proliferation of use cases. However, data quality, particularly data bias, contributes to unjust and discriminatory outcomes (Mehrabi et al., 2022). While in Contribution 2 I have explored how to make the output of ADM-based ML fair, in Contribution 4 and 5, I focus on sources of bias. In particular, I work on two different potential mechanisms that lead to unfairness in complex ML systems. In Contribution 4, I study how *privacy legislation*, *user behavior*, and *enterprise data policy* interact, solidifying algorithmic bias and unjust outcomes. This vicious cycle arises from the exhibited control that privacy legislation gives to users about their data, which, from the perspective of a single user, is warranted but collectively creates a deteriorating effect on the overall ML system's social impact. I also benchmark several technical strategies for how enterprises can adapt their ML systems such that they are more resilient against the described vicious cycle. In Contribution 5, I study how to combine fairness and the protection of privacy for tabular data used in ML for business applications. Given that it is ethically warranted to consider ethical ML more holistically in the study in Contribution 5, privacy and fairness in enterprise data management. Contribution 5 employs an extensive benchmarking framework to examine how privacy protection affects fairness. Specifically, it evaluates protection methods, such as data perturbation and synthetic data generation. The study is embedded conceptually between two key enterprise roles: data management officers who apply privacy protections and data scientists who build ML models. Data scientists require customer data to build models but cannot access this data directly. Instead, data officers must first apply privacy protection methods before releasing the data. The research seeks to understand how these privacy measures affect the downstream performance of models. Contributions 4 and 5 thus extend the overall research of this thesis about the sociotechnical implications of ML to include not only fairness and non-harmfulness but also other important pillars of responsible ML, namely privacy. As the absence of privacy also provides harm to individuals, these contributions provide another lens on the sociotechnical ML in IS. Furthermore, congruent with the instrumental objectives, the design of ADM-based systems that align with values also affects how well humans respond to such systems, correlating positively with IT adoption (Ågerfalk et al. 2022)

Based on a structure from the literature, which was presented in a deep learning context but is also applicable to other ML methods (Samtani et al., 2023), I classify the contributions in terms

of their contribution to the knowledge base in ML in IS. In Contribution 6, I apply a balanced sociotechnical lens to the development of ML-based systems. Recognizing that typical sourcing tasks are mundane and benefit from human-machine augmentation (Ågerfalk et al., 2022, p. 425), this contribution proposes a new workflow that helps organizations achieve better business outcomes (Samtani et al., 2023, pp. 13–19). In the Contribution, a prescriptive procurement solution is developed that supports procurement agents to guide their sourcing decisions, balancing direct procurement costs with indirect procurement process costs. The Contribution develops a prescriptive model that calculates expected savings from obtaining quotations from suppliers at lower prices, based on data of the current and past sourcing processes through a Bayesian learning strategy. The contribution also acknowledges a key social challenge within organizations: procurement agents in enterprises often misreport historical sourcing data to overstate their success. A key aspect is that a DSS in procurement also operates on misreported data. We evaluate the system we developed to establish rigor in the DSS design routine (Arnott et al., 2012). Overall, Contribution 6 contributes to the decision support systems (DSS) literature, especially to suggestion-models (Arnott et al., 2012), by supporting managerial decision-making (Arnott et al., 2012). Contribution 6 contributes to the ML literature in IS by the concrete application of ML in a business function, which is a common research area in IS (Abbasi et al., 2016; Padmanabhan et al., 2022; Samtani et al., 2023).

Contribution 7 sits at the more technical end of the sociotechnical spectrum. It aligns with the "learning-type" of deep learning contributions in IS (Samtani et al., 2023, p. 18). In this Contribution I investigate a learning strategy for deep learning that is based on optimizing a deep neural networks (DNN) at the weight level for multiple objectives that differ from predictive performance but optimize for business outcomes. Using the setting of Contribution 6 as an exemplary study, a DNN is optimized for direct purchase costs and process costs. The method described in Contribution 6 can be applied on a broader scale than Contribution 6, whereas Contribution 6 can also be used for a broader class of multi-objective deep learning. The contribution belongs to the predict-and-optimize paradigm (Vanderschueren et al., 2022) for modeling a business problem. To develop a precise modeling approach for the business problem, the contribution applies a heuristic strategy based on multi-objective particle swarm optimization (PSO) to train DNNs. This contribution, therefore, serves both instrumental and humanistic goals (Sarker et al., 2019), enabling organizations to explicitly articulate their objectives rather than defaulting to conventional loss functions that merely describe predictive problems. Conventional approaches like a standard loss function may fail to align with the broader social context of ADM application in certain settings.

Contribution 8 contributes to the application domain of financial trading. Contribution 8 shares heuristic optimization with Contribution 7. The research in Contribution 8 is guided by two key principles: trading decisions should be interpretable, and past stock market movements can predict future movements without individual stock predictions. To achieve this, the study employs an encoding that effectively represents market movements while maintaining rule interpretability through a hybrid learning approach combining unsupervised and supervised learning elements. This hybrid model allows for the discovery of latent market structures and the refinement of predictive accuracy based on historical data, aiming for a robust and transparent trading strategy.

Contribution 9 demonstrates a novel spatial network encoding for bike-sharing station placement optimization. The contribution introduces the conceptual framework of a "bike-sharing-station-network" and presents an encoding approach for ML in spatial networks. This encoding systematically incorporates origin locations, potential target areas, the network of reachable stations, and spatial weighting of station attractiveness based on proximate Points of Interest (POIs). The methodology aims to capture the complex spatial relationships and dependencies in urban mobility systems, advancing the representation capabilities of learning models in transportation infrastructure planning. In terms of business application, Contribution 9 can enhance operational outcomes and strategic decision-making. While numerous prior studies address initial bike-sharing system (BSS) deployment (green-field approaches), far fewer examine the ongoing operational management of established networks. The developed DSS enables the redesigning of existing networks by modeling the addition and removal of stations, allowing more targeted resource allocation for BSS operators.

Together, Contributions 6-9 make methodological contributions to ML in IS (Padmanabhan et al., 2022). The technical perspective in Contributions 6-9 is complemented by the more socio-technical perspective in Contributions 1-5. Contributions 1-5 study the "dark sides" of ADM and ways to mitigate adverse effects of AI/ML-ADM (Ågerfalk et al., 2022). Contributions 1-5 study the role of ADM in complex ML systems (Padmanabhan et al., 2022). All contributions revolve around the IT artifact conceptualized as ADM run ML/AI for automation and augmentation of decision-making.

A.7 Summaries of Papers

The first two papers form a series surrounding the question of fairness in ML specifically the empirical study of the perceived fairness of ML systems. Perceived fairness, as operationalized in this thesis, is concerned with how ML systems are perceived by humans i.e. how humans

feel about the fairness and justice of the decisions that these systems take. Certain literature streams divide perceived fairness in ML into the aspects of procedural and distributive fairness. Distributive fairness concerns the outcomes of ML decisions, whereas procedural fairness concerns characteristics of ML, such as development procedures or data usage, rather than decisions. In this thesis, I do not distinguish between these two aspects of fairness as distinct constructs. Instead, I view the distributional aspects as integral to the ML procedure, particularly through tuning of ML by suitable methods. Tuning entails modifying procedural characteristics, enabling the ML decision to become fairer and just (Friedler et al., 2019). However, the question remains how to ensure that ML systems are tuned in terms of technical parameters such that these parameters correlate with the psychological-behavioral construct of perceived fairness. Therefore, the first of the two papers conducts a structured literature review of the relevant literature on the question of optimizing perceived fairness in ML decision-making. The findings of the review suggest that there is a general agreement that the human concept of fairness encompasses a much wider range of topics than those that are covered by the horizon of existing fairness metrics. In addition, analyzing the reviewed literature, finds that no consensus has arisen with respect to how technical definitions of fairness align with human perceptions of fairness. Consequently, the second paper in this thesis is an experimental study on the topic of technical fairness metrics. In particular, in a joint effort with coauthors, I study a scenario that was previously not studied in the literature. Our findings provide a better understanding of how the human perception of fairness works for engineering ethical ML systems. To sum up, the first two papers address that these systems are behaving ethically such that their decisions correlate with societal norms of just decision-making (i.e., perceived fairness).

The behavioral component of responsible ML is just one part, as ADM-ML systems involve both human and machine components. The sub-objective RO1b addresses primarily the technical component. This sub-objective addresses questions on responsible data governance, which are a concern for organizations on how to manage their data assets about personal data (e.g., customers, employees). Organizations store this type of data in compliance with governmental or societal regulations. Nevertheless, although storing data may be seen as critical because of privacy concerns, storing data also allows the development of ML-based services and management of operations that facilitate profitability for organizations that employ ML in the first place. Data and the services built from it are therefore not only beneficial to organizations but also for customers/employees, as the services and ML applications built from it can give recommendations that give the person receiving this recommendation value.

Reiterating on the theme of fairness in ML, it is also important to understand how unfairness and algorithmic bias can emerge in ADM-ML systems in the first place. The use of privacy-preserving technology and the concept of algorithmic fairness should be mentioned here. On the one hand, algorithmic fairness aims to ensure that all social groups can benefit equally from ML technology. On the other hand, algorithmic fairness is concerned with how to avoid existing inequalities (e.g. in the form of unrepresentative data sets) or human discrimination (e.g. through the data examples that the machine uses to learn the strategy behind a human decision) being reinforced by the use of ML learning. The work in this thematic complex is also particularly concerned with how fairness interacts with privacy. The sub-objective RO1b is addressed in total by two consecutive papers.

A.7.1 Fair Engineering of Machine Learning Systems–Lessons Learned from a Literature Review (C1)

As AI algorithms become more prevalent and are used to prepare and execute decisions, the fairness of the results produced by ML systems is increasingly being debated. For ISR, two relevant questions arise: Firstly, what is known about perceived fairness that can inform the development of responsible AI/ML systems? This relates to ISR as a discipline that guides the work of information systems professionals. Secondly, from a research perspective, which phenomena need to be considered when conducting research at the intersection between technical ML and perceived fairness, in order to derive recommendations for ISR and adjacent disciplines? To understand these phenomena, the contribution begins with a business-oriented ML case study illustrating practical problems when granting loan applications from a fairness perspective. The case study was chosen to be intuitive and simple, ensuring that it effectively highlights key issues in fairness in ML. A structured literature review on the perception of algorithmic fairness was then conducted using search strings related to fairness, justice, and algorithmic decision-making in the relevant literature database. The review focused on peer-reviewed research articles from 2016 to 2021. Main focus are topics such as group-based fairness metrics, transparency, sensitive attributes, humans versus ADM, and the methodology of empirical fairness research. The literature is then analyzed with respect to what is known about key constructs of using ML for automating decision-making. In particular, the literature is analyzed for the fairness of the decision of the ML-ADM with respect to a technical fairness metric that can be used to optimize ML-ADM with regard to fairness perception. Also, procedural characteristics (e.g., ML development characteristics) and contextual factors are analyzed in order to synthesize previous findings in the empirical literature. Based on this synthesis, recommendations for further ISR research are provided.

A.7.2 Human Perceptions of Fairness: A Survey Experiment (C2)

This study examines distributive fairness and group-based metrics in a low-stakes setting using an online survey experiment. Participants were presented with four vignettes depicting different fairness definitions from the literature in a job candidate pre-screening scenario. ADM is used to pre-screen resumes and select the most qualified candidates for job interviews. This is a common use-case of ADM for supporting organizational processes. While ADM is not used for making the decision which candidate is finally hired, candidate pre-screening still significantly influences which candidates are considered for the job. The study aims to understand the most preferred fairness definition in the context of human resource (selection decisions) and contribute to the existing literature on group-based metrics. In contrast to previous literature, groups were depicted anonymously to minimize response bias. Demographic groups were represented only by the colors orange and purple. Perceived fairness was measured using a three-item scale focused on overall fairness perception. Participants were shown one algorithm at a time instead of choosing between two. This approach reflects a more realistic ADM deployment scenario and its corresponding fairness measurements. We conducted an online survey experiment using a factorial survey (vignette) design with 258 participants. Each participant evaluated four different scenarios showing the outcomes of an ADM system in a job candidate pre-screening context. The scenarios were presented as confusion matrices optimized for different fairness metrics. An ordinal mixed-effects regression model was used to analyze perceived fairness among study participants based on different fairness definitions. This statistical model was chosen for two reasons. First, to reflect the ordinal nature of fairness preferences. Second, to account for repeated measurements of participants across four vignettes.

A.7.3 AI-assisted Learning Feedback: Should Gen-AI Feedback Be Restricted to Improve Learning Success? A Pilot Study in a SQL Lecture (C3)

This study investigated the implementation of Generative Artificial Intelligence (GenAI) in educational settings, specifically examining whether restricted access to AI feedback enhances learning outcomes. The research was motivated by competing pedagogical perspectives: while AI offers ubiquitous access to individualized instruction, educational theories suggest that constant guidance may inhibit the development of deep-thinking skills. In this research, an experiment was designed to evaluate this tension, hypothesizing that faded guidance (restricted access to AI feedback) would facilitate better transfer learning compared to continuous assistance, particularly for tasks requiring deep understanding rather than mere memorization.

In a pre-registered experiment, undergraduate students in a SQL programming course were randomly assigned to two conditions while using a custom-built learning platform: a full-access group with unlimited AI feedback availability, and a partial-access group with a 90-second waiting period between feedback requests. The platform incorporated OpenAI's GPT-3.5-turbo to provide contextual feedback on students' SQL queries. After a three-week study period with identical SQL homework tasks, students completed a 30-minute pen-and-paper assessment measuring both memorization and transfer skills. Comprehensive usage data was collected including feedback request frequency, time spent on the platform, and inter-event timing.

Essentially, the study addresses not only the question if machines make humans "smarter" or "dumber" but also investigates a potential user interface design that can disrupt cognitive over-reliance on a machine by preventing cognitive processes from being completely delegated to the machine.

A.7.4 The Impact of the "Right to be Forgotten" on Algorithmic Fairness (C4)

In practice, deploying ML often involves a dynamic interaction between machines and humans. ML algorithms provide recommendations, which humans interpret and act upon, demonstrating their agency over the algorithms. Agency over ML is supported by data control rights, such as the 'right to be forgotten'. For instance, when a user requests data deletion, their data is removed from the ML system. Consequently, the bilateral relationship between humans and machines shapes the behavior of these complex systems.

The contribution examines feedback loops that contribute to unfairness in ML predictions and explores how data deletion impacts predictive performance and fairness. Legal frameworks like the EU's GDPR (Article 17) empower users to control their data by withdrawing consent. Anonymization, as an alternative to deletion, removes personally identifiable information while retaining some analytical utility.

Four main technical approaches are identified that organizations can use:

- *Record deletion*: Completely removing the record from the dataset
- *Record masking*: Anonymizing only specific records through techniques like using dummy values
- *Table anonymization*: Proactively anonymizing the entire dataset
- *Attribute deletion*: Removing sensitive features entirely from the analysis.

Using the German credit dataset, the study simulated scenarios where users requested data deletion. Two scenarios were analyzed: “low” (5% of group members requested deletion) and “high” (20% requested deletion). The protected attribute was age, comparing younger and older credit applicants. The Bradley-Terry model is used to evaluate the performance of different personal data protection strategies (PDDR) strategies in terms of fairness and accuracy.

A.7.5 Balancing Privacy, Fairness, and Utility in Data Sharing: Synthetic vs. Perturbative Approaches a Robust Assessment (C5)

The increase in data collection across domains highlights the need to protect sensitive information while enabling data scientists to develop effective ML models. This study investigates how organizations can develop responsible data sharing pipelines that balance privacy concerns with the utility of data for decision-making purposes and examines the subsequent impact on the fairness of ML models. Since privacy protection methods tend to degrade model performance, it is important to consider whether this loss of performance is distributed evenly. This contribution investigates the interplay between privacy protection and fairness in ML applications, specifically in the context of tabular data used for business applications. The study addresses the challenge of balancing privacy concerns with the need for fairness in ML models, particularly when sensitive attributes are involved. The research focuses on two main aspects: the impact of privacy protection methods on fairness, and the trade-offs between privacy and fairness in ML applications. A comprehensive benchmarking framework is employed to evaluate the effects of various privacy protection methods, including data perturbation and synthetic data generation, on fairness metrics. Unlike C3, which focuses on the impact of data deletion on fairness, this study examines the effect of various privacy protection methods on fairness in ML models. The research is embedded in the context of enterprise data management, where data officers apply privacy protection methods before releasing data to data scientists for ML model development. The question is how these privacy measures impact the performance of models downstream, particularly in terms of fairness.

A.7.6 Prescriptive Analytics in Procurement Reducing Process Costs (C6)

This contribution addresses a critical challenge in procurement: balancing direct purchasing costs with indirect supplier search costs when sourcing suppliers. Supplier search costs and other related costs are subsumed under the process costs of procurement. In supplier sourcing, determining when to stop searching for better prices is challenging because obtaining price quotations requires time and resources, yet the potential savings remain unknown until these costs are incurred. The contribution frames procurement as an “exploration vs. exploitation”

trade-off. While exploring more suppliers might yield lower prices, the process costs may outweigh potential savings, especially for low-cost items like electronic resistors. Traditional approaches often use arbitrary supplier limits, which may not be optimal.

The developed prescriptive analytics artifact predicts whether searching for further suppliers is worthwhile, as the expected savings in direct purchase costs may be marginal compared to the procurement costs: finding and qualifying a supplier. Two approaches are proposed: The first approach uses only historical quotation data to guide the current sourcing process, recommending whether to continue or stop searching. The second approach, based on a Bayesian setup, can learn from the current sourcing process. The proposed approaches are compared against benchmark strategies. Several robustness tests are applied. To evaluate the relative effectiveness of proposed approaches for handling inaccurate learning data, systematic scenarios about the database's structure are evaluated. This is because the historical data used for learning may systematically differ from the actual data-generating process. Procurement agents are often financially rewarded for reported cost savings, which incentivizes them to manipulate the historical database by over-representing expensive quotes. Furthermore, an ablation study is conducted to evaluate distributional assumptions underlying the prescriptive solutions, comparing the performance on simulated data, where the true distribution is known, with the performance on the real quotation data. The real quotation data comes from an industrial dataset of procurement episodes of electronic resistors (201,187 quotations), where indirect sourcing costs are substantial compared to uncertain direct cost savings.

A.7.7 A Multi-Objective Particle Swarm Optimization Framework for Operations Management (C7)

This contribution addresses a fundamental challenge in operations management: effectively combining predictive modeling with optimization. While ML applications in operations management are growing, traditional approaches often follow a "predict first, optimize later" paradigm that fails to fully address complex business objectives.

I propose a framework based on evolutionary computing with multi-objective particle swarm optimization (MO-PSO), which designs the fitness function according to the business operations. This approach enables optimization for problems that would otherwise be difficult to solve with classical supervised learning. The contribution identifies four common challenges with supervised learning in operations management contexts, which the proposed framework seeks to overcome: First, when there is no natural quantity to predict, supervised ML methods lack labeled data for training. Second, when prediction errors are costly and imprecisely

measurable, approaches such as instance-based or class-based weights cannot accurately approximate the prediction error. Third, when the loss function is ill-defined, it does not capture the underlying optimization problem. And fourth, when there are multiple objectives to optimize, traditional optimization routines in ML cannot be used. Traditional gradient-based deep neural network training often requires compromises when faced with these challenges, while the proposed gradient-free optimization method overcomes these limitations by optimizing directly for business outcomes.

The proposed framework is illustrated by a procurement case study that focuses on sourcing suppliers and procuring items. This process involves balancing two conflicting objectives: minimizing direct purchase costs (which requires evaluating many suppliers) and minimizing procurement process costs (which requires keeping the supplier pool small). The framework uses a deep neural network to predict a "stop price"-the threshold at which the search for additional supplier bids should stop. Rather than training with traditional supervised learning, the network's weights are adjusted using MO-PSO according to the model`s operational impact. This approach directly models the consequences of predictions and optimizes for actual business objectives without requiring explicit labels or predefined loss functions, overcoming the limitations of traditional supervised learning in this context.

A.7.8 Robo-Advisory and Algorithmic Trading via Evolutionary Discretization and Rule-Mining (C8)

Financial time series are rich in information that can be extracted for insight, yet they require special pre-processing in order to be analyzed by ML algorithms. Financial time series often have a multidimensional structure, where future movements are influenced not only by their trajectory but also by the movements of related financial time series. Furthermore, interpretability is frequently desired by practitioners.

To address the above challenges of multidimensional data structure and interpretability, the study investigates a hybrid approach to financial time series analysis that combines feature engineering, an evolutionary strategy for data discretization, and rule-mining technique. Together, these techniques form interpretable algorithmic trading systems. The study addresses three key challenges in robo-advisory: detecting events in financial data streams, balancing prediction accuracy with interpretability, and generating actionable advice. An evolutionary strategy is used to optimize discretization thresholds for financial time series, transforming continuous data into discrete categories (rising, falling, steady) that enable subsequent rule-

mining algorithms to extract trading rules. The discretization thresholds are optimized using the profitability of the obtained rules.

The methodology involves pre-processing multidimensional financial time series data through normalization, feature engineering, and labeling. The trading system uses a multi-horizon framework with prediction, planning, and execution components. It allows for both short and long positions to maximize performance. Overall, the study makes contributions in developing interpretable systems that detect profitable events in time series.

A.7.9 Bike-Sharing Station Placement: Spatial Analysis and Data Mining of Network Design Characteristics (C9)

This study investigates data-driven strategies for planning BSS station locations, uniquely considering both competition from nearby stations and complementary influences within a bike trip's target area. In this research, over eight million booking records from German bike-sharing providers are analyzed to develop methods for optimizing station placement and reorganization to maximize customer utilization. The central research question addresses how changes in total daily rental time can be accurately estimated when adding or removing stations while accounting for network effects. The study focuses on station-based systems, which offer advantages over free-floating systems through smaller fleet sizes, lower maintenance costs, and easier prevention of misuse.

Spatial modeling approaches are used, which incorporate spatial dependence between stations through weight matrices. This methodology recognizes that the success of one station can influence nearby stations, either positively (complementary effect) or negatively (competition effect). To accurately assess the proposed solution, a spatial cross-validation strategy is used. The results reveal significant spatial dependencies in the BSS networks. Furthermore, an interactive visual planning tool has been developed that allows BSS planners to simulate adding or removing stations and visualize the predicted effects across the entire network. This tool utilizes an R backend with Shiny for interactive dashboards and Leaflet for map displays, fetching data from Google Maps API and sociodemographic geodatabases to make predictions.

Overall, the study contributes valuable insights for BSS operators making strategic network design decisions. By quantifying both direct effects (on the new station itself) and indirect network effects (on nearby existing stations), the methodology provides a comprehensive decision framework for network expansion planning. The results emphasize the importance of considering spatial dependencies.

A.8 Adjustments and Modifications

The following changes were made to these publications to improve readability:

- The contributions were restructured to follow a consistent format.
- The reference style was harmonized to use the APA style.
- The language was harmonized to use American English.
- The author adjusts some minor issues (e.g., grammar, etc.) in all publications.
- Terminology was harmonized (e.g., multi-objective and multiobjective)
- Enhanced C7 with a more comprehensive weight initializer description and updated with accurate citation references.
- Consistent mathematical notation.
- References are at the end of the thesis instead of at the end of each individual contribution.
- Inserted footnote Nr. 6 after publication for clarification of concepts.
- Table 1.3 inserted text in first header cell which was empty in original submission.
- The term Algorithm was replaced by Listing.
- Figure 2.1 was remade in a higher resolution to improve depiction quality.
- A previously missing differential was added on page 141.

B: Contributions

"All sorts of things can happen when you're open to new ideas and playing around with things."

Stephanie Kwolek

I. Ethical Foundations of Machine Learning Systems

Humans have an innate understanding of fairness from an early age. Nevertheless, while humans are able to understand fairness intuitively, or the absence of it, it is unclear how to engineer this sense of fairness into a technical concept. This is especially true for machine learning systems, which are often seen as ‘black boxes’ that produce results without clear explanations. As a result, it is important to understand how humans perceive fairness and how this perception can be translated into technical definitions of fairness, which can be used to detect unfair decision-making of an ML system and then use it as a criterion to improve the performance of that system on that criterion. In this chapter, we will explore the concept of fairness in machine learning systems and how it can be operationalized. Congruent to make machine learning systems fair, I also explore specific emerging risks of GenAI technology when used in education.

1. Fair Engineering of Machine Learning Systems – Lessons Learned From a Literature Review

Julian Sengewald (TU Dortmund)
Richard Lackes (TU Dortmund)

Abstract. With the growing prevalence of AI algorithms and their use to prepare and even execute decisions, there is increasing debate about whether the results of machine learning systems tend to be fairer or more unfair. When faced with engineering a fair machine learning solution in practice, trade-offs arise between conflicting fairness notions. We conduct a literature review on this topic. The results of our review indicate that a slight consensus exists that the human concept of fairness is much broader than what lies in the scope of current fairness metrics. We discuss the context of judging fairness metrics. We also find that, albeit much research already has been done, there is room for improvement when seeking to generalize the findings across different scenarios.

1.1 Introduction

Because of documented misbehavior in machine learning algorithms, the topic of algorithmic fairness has attracted much attention in recent years. For example, in healthcare applications (Obermeyer et al., 2019), crime prediction (Angwin et al., 2016), or ad-delivery (Latanya Sweeney, 2013). All these cases have raised a significant debate about algorithmic fairness in research. For example, research was conducted on synthesizing the causes of unfairness in machine learning (Mehrabi et al., 2022), algorithmic measurement of fairness (Herington, 2020), or optimization methods to achieve a certain notion of algorithmic fairness (Haas, 2019).

1.2 Background

1.2.1 Machine Learning

Firstly, machine learning systems are specified to define fairness criteria precisely. The machine learning system $h(\cdot)$ allocates a benefit to an individual instance x if $h(x) = 1$. The true class label y provides additional information, where $y(x) = 0$ indicates ineligibility and $y(x) = 1$ indicates eligibility. If the system predicts $h(x) = 1$ but $y(x) = 0$, it results in a false-positive (FP), whereas $h(x) = 0$ but $y(x) = 1$ results in a false-negative (FN). In all other cases, the prediction is correct. Often, predictions made by the machine learning system are

defined on a probability domain, i.e., $h(x) \in (0,1)$, which can be interpreted as a score. Instances are classified as eligible if the score exceeds a predefined threshold, i.e., $h(x) > \tau$.

1.2.2 Gateway and Selection Decisions

We classify two types of decision-making that one can find in problems where machine learning may be applied, and fairness is a concern: (1) gateway decisions (2) selection decisions. A *gateway decision* is characterized by having to decide about the treatment of a particular instance. Depending on the decision at the gateway, the instance would experience completely different treatments (e.g., bail or no-bail decision). In such applications, we are primarily concerned with the quality and the costs and harms of a wrong decision. The costs of an FP and FN are determined by the wrong submission to a certain branch of a treatment process. *Selection problems* are due to limited resources, such that even when $h(x) = 1$ not every instance receives the benefit (e.g., resume selection for job interviews). If there were infinite resources, there would be no classification costs (and no selection problem). Thus, in selection problems, the cost of an FP is mainly defined by the fact that an FN cannot receive the benefit. Distinguishing between those two types of decision problems may help to understand situations of unfairness.

1.2.3 Fairness Metrics

Fairness can be defined either at the individual level or at the group level (Hutchinson & Mitchell, 2019). We are concerned with fairness at the group level. Technical fairness measures can quantify systematic biases in machine learning systems that lead to disproportionately harming one group. Different fairness models can be defined with this configuration:

- Demographic parity (also known as statistical parity): equal allocation of the benefit, e.g. (Feldman et al., 2015)
- Equalized odds: equal true-positive-rate and equal false-positive rate across groups (Hardt et al., 2016)
- Equal opportunity requires an equal true-positive-rate across groups (Hardt et al., 2016)

The precise mathematical definitions of these metrics and their components are given in Table 1.6. An example of the meaning of a fairness metric is $P(h(x) = 1 | \, x \in g_i)$, which is the probability of how often the machine learning system will allocate a benefit to the group i.

1.3 Case study

The case study illustrates practical problems when engineering a fair machine learning solution, which motivated the following literature review. We used the German credit dataset from the UCI machine learning repository for the case study. The task of the case study is to predict failed/non-failed credits according to a set of input attributes. This problem is modeled using logistic regression. Suppose there is only enough capacity ϕ at the bank to process 100 credits. The bank would decrease the score predicted by the machine learning model and grant credits to the 100 top applicants. The resulting threshold is then τ (e.g., 0.93). Suppose that there are two groups $g_1 = \{x: x_{Age} \leq 25\}$ and $g_2 = \{x: x_{Age} > 25\}$. The corresponding fairness metrics are given in Table 1.1.

Table 1.1: Equal group threshold

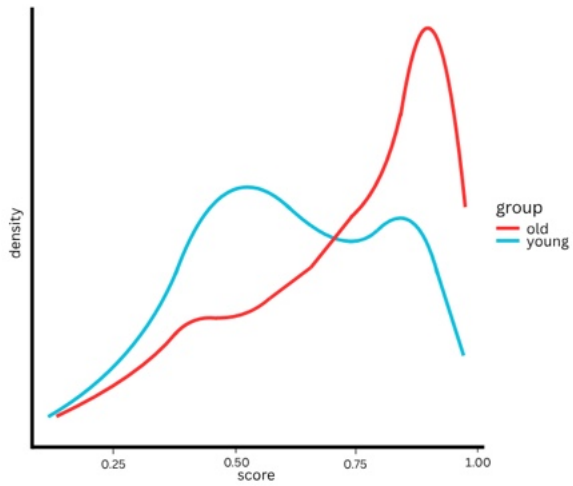
Fairness metric	g_1	g_2
$P(h(x) = 1 \mid x \in g_i)$	0.02	0.12
$P(h(x) = 1 \mid y(x) = 1, x \in g_i)$	0.04	0.16
$P(h(x) = 1 \mid y(x) = 0, x \in g_i)$	0.00	0.02

The result of this procedure is unfair for the younger group of credit applicants. The reason for this lies in the distribution of the score across groups (Figure 1.1). The distribution of scores for younger people is shifted to the left. A problem here is, of course, that the loan default yes/no plays a role as well as the loan amount. On average, the defaulted loans of younger borrowers amount to only 82% of the loan amounts defaulted by older borrowers.

Lower defaulting debt for younger borrowers could justify a group-specific threshold, as the financial risk for the bank is lower for younger individuals (assuming repayment and interest rates are not considered). Equal access to financing is a societal concern because individuals may not have had sufficient opportunity opportunities to develop the financial strength needed to pass a credit application check, often leaving them excluded. Addressing this issue, we set a group-specific threshold τ_g' where we spread the resources proportionally according to a convex combination of the group "eligibility"-rate $\phi * P(x \in g_i | y(x) = 1)$. Note that this also meant that $\tau_2' > \tau$ and $\tau_1' < \tau$. The corresponding fairness metrics are reported in Table 1.2. While the situation for the young has improved, the older are now slightly worse off than before.

A third suggestion would be to skip the age attribute from the machine learning model. Indeed, the overall fairness situation improves, but overall performance goes down because more

Figure 1.1: Distribution of scores across groups



ultimately defaulting credits will be classified as non-defaulting credits. Similar considerations can also be done for different machine learning algorithms, yielding different fairness and overall predictive performance.

Since there the fairness metrics pose a trade-off, and we ask which fairness metric does correlate most with layman perception of fairness. We formulate our research questions (RQ):

RQ1: *What is the current state on which fairness metrics will be regarded as the fairest by the public?*

RQ2: *What contextual factors are important for implementing human notions of fairness into fairness metrics?*

To investigate this subject, we conducted a review of the related literature.

Table 1.2: Unequal group threshold

Fairness metric	g_1	g_2
$P(h(x) = 1 \mid x \in g_i)$	0.06	0.11
$P(h(x) = 1 \mid y(x) = 1, x \in g_i)$	0.09	0.14
$P(h(x) = 1 \mid y(x) = 0, x \in g_i)$	0.01	0.02

1.4 Research Methodology

1.4.1 Search Query

We employ a variety of compositions of search strings [(“fairness” OR “justice”) AND (“judg*” OR “perce*”) AND (“machine learning” OR “artificial intelligence” OR “algorithmic decision-making”)]. The search was conducted for the database fields abstract and title.

1.4.2 Time Period and Other Search Criteria

We chose the years from 2016 to 2021 as the time period for this research. This is because the topic of algorithmic decision-making would not have been generally understood in the general public population. If the database offered the option only to include peer-reviewed research, we chose that option; otherwise, we used peer review as an inclusion criterion for the search results.

1.4.3 Inclusion Criteria

As inclusion criteria for all search results from the primary search query, we choose:

- The title, abstract, or introduction of a paper must be related to the perception of algorithmic fairness.
- The paper is a peer-reviewed research article (including conference proceedings).
- The paper is about empirical research, not a technological artifact, algorithm, method, or philosophical discussion.

Table 1.3: Databases searched

Database\Result	Found	Included	Backward	Forward	Total
ACM	17	11	4	8	23
IEEE/AIS	0				
Total	17	11	4	8	23

The databases we have chosen reflect the associations related to the community of information systems.

1.5 Results

1.5.1 Overview

We identify the following meta topics in the literature as summarized in Table 1.4:

Table 1.4: Meta topics in literature

Meta topics	References
Fairness Metrics	(Cheng et al., 2021; Harrison et al., 2020; Mohammad et al., 2019; Saxena et al., 2020; Srivastava et al., 2019; Wang et al., 2020)
Transparency	(Binns et al., 2018; Dodge et al., 2019; Kizilcec, 2016; Schöffer et al., 2021)
Use of sensitive attributes	(Albach & Wright, 2021; Berkel et al., 2019; Grgić-Hlača et al., 2012; Schöffer et al., 2021)
Human vs. ADM	(Araujo et al., 2020; Harrison et al., 2020; Langer et al., 2019; M. K. Lee, 2018; M. K. Lee & Baykal, 2017; Lee & Rich, 2021; Marcinkowski et al., 2020; Schöffer et al., 2021; Wang et al., 2020)
Methodology	(Georg et al., 2021; Hannan et al., 2021)

The studies also differed in the scenarios that were considered. However, most studies dealt with problems in legal justice (esp. risk of reoffence prediction), as shown in Table 1.4.

Most studies considered gateway decisions; only a few studies concerned resource allocation and selection: M. K. Lee & Baykal (2017).

In the following, we present more detailed results:

Table 1.5: Overview of scenarios

Work	
Hiring (resume selection)	(Langer et al., 2019; Lee, 2018)
Evaluation, promotion	(Binns et al., 2018; Lee, 2018)
Task scheduling	(Lee, 2018)
Training	(Langer et al., 2019; Wang et al., 2020)
Justice	
Small offences (parking tickets)	(Araujo et al., 2020)
Starting prosecution/lawsuit	(Araujo et al., 2020)
Risk of reoffence (Granting parole/bail)	(Albach & Wright, 2021; Berkel et al., 2019, 2021; Harrison et al., 2020; Mohammad et al., 2019; Srivastava et al., 2019)
Child protection	(Albach & Wright, 2021; Cheng et al., 2021)
Education	(Georg et al., 2021; Marcinkowski et al., 2020)
Health	
Diagnosis and Treatment	(Albach & Wright, 2021; Araujo et al., 2020; Georg et al., 2021; Lee & Rich, 2021; Srivastava et al., 2019)
Fitness recommendations	(Araujo et al., 2020)
Autonomous Driving	(Awad et al., 2018)
Media (News recommendation)	(Araujo et al., 2020)
Account Blocking (Banking, Social Platforms)	(Araujo et al., 2020; Binns et al., 2018)
Banking Loan	(Albach & Wright, 2021; Binns et al., 2018; Saxena et al., 2020)
Social Welfare	(Albach & Wright, 2021; Georg et al., 2021)

None/not classified	(Helberger et al., 2020; Woodruff et al., 2018) / (Lee & Baykal, 2017)

1.5.2 Fairness Metrics

First, we matched the fairness metrics in each study to the closest fairness metric according to our classification. The aggregated result is depicted in Table 1.6.

Table 1.6: Overview of fairness metrics

Fairness Metric	Frequency
Demographic Parity (=DP) $P(h(x) = 1 \mid x \in g_i) = \vartheta$	4 (Cheng et al., 2021; Harrison et al., 2020; Saxena et al., 2020; Srivastava et al., 2019)
Equal error rates (EER) $P(h(x) = 1 \mid y = 0,\ x \in g_i) = \vartheta_{error,\,0}$ $P(h(x) = 0 \mid y = 1,\ x \in g_i) = \vartheta_{error,\,1}$	4 (Cheng et al., 2021; Harrison et al., 2020; Srivastava et al., 2019; Wang et al., 2020)
Equal false-positive-parity (=FPP) $P(h(x) = 1 \mid y = 0,\ x \in g_i) = \vartheta_0$	1 (Srivastava et al., 2019)
False-negative-parity (FNP) $P(h(x) = 0 \mid y = 1,\ x \in g_i) = \vartheta_1$	1 (Srivastava et al., 2019)
Equalized odds (ED)	2 (Cheng et al., 2021; Saxena et al., 2020)
Equal accuracy (=Acc) *Equal error rates (EER) imply equal accuracy*	4

The matching of the studies to the corresponding fairness metrics was carried out as depicted in Table 2.6. We matched demographic parity with the following denominations in the papers =DP (Srivastava et al., 2019), equal outcomes (Harrison et al., 2020), equal resource allocation (Saxena et al., 2020), statistical parity (Cheng et al., 2021). We matched EER with the following definitions in the corresponding papers: equal error rates (Wang et al., 2020), EP (Srivastava et al., 2019), equalized odds (FPP and FNP) (Cheng et al., 2021). The term accuracy was used in (Harrison et al., 2020). EER imply equal accuracy and vice versa; hence, this implies equal

accuracy if one favors EER. Therefore, the matching of equal accuracy and EER is the same. FPP and FNP, we matched FNP and FPP (Saxena et al., 2020). For equalized odds, we matched the equalized odds from (Cheng et al., 2021) and equal rates from (Saxena et al., 2020). The latter considers a decision in which a limited number of resources is split proportionally to repayment ability (Saxena et al., 2020). This implies that the decision is independent of group membership but conditioned on the true outcome and. In our view, this closely matched the equal opportunity and equalized odd definition of fairness, which measure meritocratic and non- meritocratic allocation. Although it should be bearded that (Saxena et al., 2020) studied a problem for allocating a continuous benefit, equalized odds were initially proposed for binary outcomes. Further, they consider individualized instead of group fairness (Saxena et al., 2020). It may be that fairness at the individual level is considered differently than at the group level; but this may also depend on the amount of information given. The way in which the study is conducted, there was no difference between the two compared individuals except their ethnicity (group membership) and repayment ability. Concerning our research questions, we looked at each study, comparing pairwise the metrics under consideration and counted how often they performed better (i.e., preferred by a higher number of people) against the remaining metrics. DP was in 57% (4/7) pairwise comparison the most preferred metric (Cheng et al., 2021; Harrison et al., 2020; Saxena et al., 2020; Srivastava et al., 2019). ERR was in 33% of the pairwise comparisons (3/9), the most preferred metric (Harrison et al., 2020; Srivastava et al., 2019). FPP was in 83% (5/6) cases the most preferred metric (Harrison et al. 2020). FNP was in 25% (1/4) of the cases preferred (Srivastava et al., 2019), but this could be due to the framing (see also Section 4.3). Finally, considering the comparison of equalized odds with DP, the former was always preferred (Cheng et al., 2021; Saxena et al., 2020). One study contained metrics that we did not find in other studies (Srivastava et al., 2019), and they were also the least preferred; those metrics were excluded. We also checked the qualitative results of the studies we reviewed. In qualitative interviews, experimental subjects were not always willing to sacrifice overall accuracy for increased fairness (exceptions include if a larger or more disadvantaged group would benefit) (Cheng et al., 2021). A slight preference for favoring the disadvantaged group (affirmative action) was also found in other contexts (Saxena et al., 2020). Hence, one should also take such aspects as group size and disadvantage level into account. Such aspects may be helpful when developing new fairness metrics. How can these results be interpreted? First, ED and FPP appear in conjunction to be the most favored metrics. The next most favored is DP. Albeit DP would imply that we would not have a decision and thus a machine learning problem. So, one needs to be wary of over-interpreting that result. Also, the

way of aggregation can affect the ranking. We aggregated over pairwise comparison within a separate analysis in the literature we examined. A caveat is that this could give a single study much weight if this study conducted many comparisons.

1.5.3 Experimental Procedures

After having reviewed the experimental procedures, we identified four criteria that could contribute to the further comparability and futility of such studies in the future:

- Framing costs of wrong decisions.
- Visualization of scenarios.
- Availability of no-choice option.
- Defining the target population

The framing of the costs of a wrong decision might explain two seemingly contradictory results concerning the preference for equalizing FP. For instance, in (Srivastava et al., 2019) the costs of FP and FN were disproportionate in the two scenarios examined (granting parole vs. diagnostic analysis in healthcare). The cost of a FP and FN are difficult to compare when predicting the risk of reoffence for the purpose of granting parole, because in the case of a FP, the cost of inaccuracy is borne by the convict, but the cost of an FN is wholly borne by the society. The description of the scenario given *"[…]A defendant falsely predicted to reoffend can unjustly face longer sentences, while a defendant falsely predicted not to reoffend may commit a crime that was preventable"* (Srivastava et al., 2019) also (over-)emphasized this aspect in comparison to the case of health risk prediction where *"[…] a patient falsely diagnosed with high-risk of cancer may unnecessarily undergo high-risk and costly medical treatments, while a patient falsely labeled as low-risk for cancer may face a lower chance of survival"* (Srivastava et al., 2019). We expect non-medical specialists to struggle to balance FN and FP costs in the healthcare scenario, whereas in the scenario related to crime, they may regard FN worse than FP as they may be affected firsthand by an FN. Other research related to machine learning applied to justice administration sought to frame FN and FP more comparable by considering bail/no-bail decisions for non-violent crimes and not mentioning the possibility of committing further crimes for FN (Harrison et al., 2020). Differences may exist due to the samples' differing compositions. Healthcare scenarios are also sensitive towards institutional cross-country differences, such as the existence/coverage of public insurances. Apart from institutional differences, there are also cross-cultural differences (Awad et al., 2018). However, most studies have been conducted with samples obtained from Amazon Mechanical Turk (MTurk) possessing a similar composition in the studies under consideration.

Visualization of scenarios might be another point to consider when designing experiments. Visualizations in the reviewed literature can be divided into instance-based and aggregated depictions. Instance-based representation can be, for example, binary, in which a single instance received or did not receive the benefit (Wang et al., 2020), or pairwise, in which the classification of two individuals are compared (Saxena et al., 2020), or depict multiple instances (Srivastava et al., 2019), or be supported by pictorial depiction (Cheng et al., 2021; Srivastava et al., 2019), including the use of confusion matrices (Cheng et al., 2021). Aggregated depiction can be based on a multi-metric (Harrison et al., 2020) or supported by a diagram (Harrison et al., 2020). In sum, many visualization types have been used in experimental studies. Some researchers, however, pointed out that the cognitive load incurred by complicated visualization practices, e.g., multiple instances, could affect comparability because experimental subjects do not fully grasp the actual situation (Harrison et al., 2020). Detailed pictorial depictions of multi-instance situations require the experimental subjects to mentally calculate fairness metrics, whereas, in single/multi-metric representations, the aggregation has already been done.

Furthermore, photographs, when presenting the experimental vignettes, influenced how female experimental subjects judged fairness (Mallari et al., 2020). This is actually of importance because of the plenty of results about how perceived fairness can be affected by demographics (e.g., age), and the domain of decision can also affect the effect of demographics (Georg et al., 2021) At the same time, the demographics of the experimental participants themselves did not affect the unfairness perception of using a demographic attribute (Berkel et al., 2019). Another question is which visual best aids comprehension. First, when using a confusion matrix, it seems better to use a contextualized one (i.e., giving actual names to positive and negative outcomes) (Shen et al., 2020). Second, for the task of comparing situations, contextualized confusion matrices are understood as well as bar charts (Shen et al., 2020). Visualization improves experimental participants' comprehension but employing pictures should be carefully considered as it impacts the ratings.

We discovered only one study that included a no-choice option (Srivastava et al., 2019); the majority of studies forced a choice between a predefined list of fairness metrics. On the one hand, it makes sense from a practical standpoint to evaluate "established" fairness metrics and then choose the one that best correlates with the human judgment of fairness. But, since we do not (yet) know which fairness metric is most suitable for human perceptions of fairness, one cannot know beforehand if the list of fairness metrics is exhaustive. This is a striking point because, in the study that included the "no option preferred"/"do not know" category, it was a relatively often chosen category (Srivastava et al., 2019). For example, research on survey

methodology points out that the "do not know" option for attitudes is attributed to ambivalence and ambiguity (Krosnick & Presser, 2010). Hence, the omission of the "no option preferred"/"do not know" category could seriously affect results. Nonetheless, careful consideration is required because "do not know" was also found increasing satisficing behavior (Krosnick & Presser, 2010).

All studies either measured fairness preference elicitation either by using a Likert scale, e.g., (Harrison et al., 2020; Wang et al., 2020), or binary choice between two alternatives, e.g. (Srivastava et al., 2019). In terms of responses, there seems no substantial difference between the two measurement scales (Georg et al., 2021).

A general question is if crowdsourcing of fairness perceptions is desirable. First, many research studies use platforms like AMT and obtain skewed samples of the general population (e.g., age Albach & Wright, 2021; Srivastava et al., 2019). Nevertheless, the impact of such samples can be reduced statistically (Chen et al., 2023) or with in-experiment stimuli. Such in-experiment stimuli can be used before conducting the fairness perception measurement by providing deliberate exposure to varied viewpoints and diversity; this shifts a small group's vote to more closely representing the majority vote (Berkel et al., 2019, 2021). Hence, crowdsourcing results can be applied to a larger population, given that one stimulates diversity in thinking. However, prior literature also points out that algorithmic fairness is all about that algorithms, machine learning, and AI work well for minorities and disadvantaged groups in society (Mohammad et al., 2019). It may be questioned if majority votes are the best course of action for future research on fairness perceptions. Henceforth, more research into the perceptions of fairness among marginalized populations may be critical, e.g. (Lee & Rich, 2021). For example, we may employ student populations because they are more likely to understand the implications of algorithms used in the recruitment process as they are affected firsthand. Though, one must remember that students make up most prospective hires and that people still change careers at a later age. Prior studies of in-sample differences in demographic effects are also mixed (Albach & Wright, 2021). Hence, an important question is who the target population is when doing crowdsourced design of ethical AI systems.

1.5.4 The Context of Judging Fairness Metrics

We discussed the fairness perceptions of fairness metrics in the preceding sections, but the context of fairness judgment also needs to be considered for understanding the limits and potential future avenues for research. Since all experiments usually involve asking an

experimental subject to judge a situation affecting a group of individuals, it seems natural to consider the contextual effects of who you ask to judge whose allocation.

First, the recipient's attributes should be considered (e.g., age, gender, or ethnicity). Overall, some demographic attributes such as age (Berkel et al., 2019; Cheng et al., 2021; Georg et al., 2021), health status (Georg et al., 2021), criminal history (Georg et al., 2021), having children (Georg et al., 2021) seem more fair or acceptable for model inclusion than ethnicity and gender (Berkel et al., 2019; Cheng et al., 2021). This can be situation-specific but primarily not dependent on the relationship between the recipient and the fairness judger (except political partisanship) (Georg et al., 2021).

Prior experience with AI-based ADM increases its fairness perception (Schöffer et al., 2021). In division tasks, the subject's outcome compared to the outcome is seen to be less fair as compared to human decision-making as a group; the more the subject knows about the algorithm (computer programming) and the greater their interpersonal power (Lee & Baykal, 2017). The latter aligns with another finding that revealed that mathematics and natural science majors were less inclined towards protesting against ADM (Marcinkowski et al., 2020). These findings imply that education in machine learning and ADM makes humans believe more in technology. This is not necessarily a good thing, given the documented AI misbehavior (Köchling & Wehner, 2020; Sweeney, 2013) and that ML educated are also the ones that typically are the ones that apply ML.

On the other hand, self-perceived marginalization reduces the fairness perception (Wang et al., 2020). In addition, differences depend on prior expectations concerning the outcome. Individuals who do not receive a benefit allocated by ADM but anticipate qualifying for it have a more pronounced perception of unfairness (Wang et al., 2020). The management of (unwarranted) expectations seems therefore also crucial in ADM. Also, prior distrust in the domain where ADM is deployed may reduce fairness perception of ADM in comparison to human decision-making (Lee & Rich, 2021). Hence, that is a similar phenomenon as self-perceived marginalization, which can reduce fairness perceptions (Wang et al., 2020).

Another approach to fairness is considering several dimensions at the same time. Such a multidimensional study was proposed to evaluate characteristics of a person (circumstances) that should not affect the amount of benefit (utility) they receive given the same level of meritocracy (Albach & Wright, 2021). This augments the ED metric to include affiliation with multiple groups and an individualized utility quantification of the received benefit. They found augmented ED increases utility perceptions. Another vein of literature looked at the used

features and studies if properties of these features explain fairness judgment (Albach & Wright, 2021). Interestingly, a set of features collectively is predictive of fairness perception of ML decision outcomes across situations, suggesting that feature properties explain situational fairness perception (Albach & Wright, 2021). Secondly, looking at the single most predictive property, humans mainly evaluate the relevance and truthfulness of a feature (Albach & Wright, 2021). Other literature found unrelated demographic attributes were not acceptable, even if they increased accuracy (Grgić-Hlača et al., 2012). This suggests that the situational relevance of demographic attributes is critical for experiments on the perception of fairness, including those on metrics.

A contextual factor affecting the ratings of the fairness metrics could be the domain in which ADM is applied. Human decisions are considered fairer even in tasks that usually require human skills and allow for human biases (work assignment/scheduling, hiring, work evaluation), as found in survey experiments (Lee, 2018). Qualitative results from the previous study hint that humans perceive human decision-making as more fair because it may consider nuanced contextual factors (e.g., holiday plans) and be less sensitive to errors (Lee, 2018). Similar results were also obtained in laboratory experiments on division tasks (e.g., sharing rent, good division) where ADM was perceived as less fair than human decisions, where a group discussion achieved the latter. It was noted that humans' perception of fairness is often rather holistic and comprises altruism/pro-social behavior. However, the capability of holistic perspective-taking was not attributed to ADM, possibly explaining why they were perceived as not fair (Lee & Baykal, 2017). ADM is also perceived as less fair if ADM is done too extensively (as compared to partial ADM) (Langer et al., 2019) or done in high stake situations such as criminal justice (Araujo et al., 2020; Harrison et al., 2020) and resume screening (Langer et al., 2019).

In contrast, for the scenario of university admission, ADM was perceived to be fairer than human decision-making (Marcinkowski et al., 2020). This could be because the decision-making attributes employed were perceived as relevant properties (Albach & Wright, 2021). Similarly, ADM was preferred when asked about the general fairness of ADM vs. human decision-making (Helberger et al., 2020). Those studies were non-MTurk samples conducted in Germany and Netherlands (Helberger et al., 2020; Marcinkowski et al., 2020). Another factor for ADM's perceived fairness here may be that algorithmic unfairness problems have received less attention in Europe than in the US, where most research has been done.

Research on other than the before-mentioned scenarios did not find a significant difference between humans and ADM in health-related issues and the media (Araujo et al., 2020). An explanation may be that those scenarios are inherently different from work-related scenarios, or that the scenario description was overall shorter (4-10 words, Araujo et al., 2020) than in other studies (33-76 words (Lee, 2018), or the experiment was not facilitated by support staff (Lee & Baykal, 2017).

Concerning the experimental procedures, we noted there are also some problems with scenario settings. As revealed by answers to open-ended questions, humans could misinterpret judging the fairness of the overall situation instead of comparing the outcome as produced by either a machine (e.g., judging short notice, or that a process is fair because everybody is subject to the process, or the transparency of the process) (Lee, 2018).

In sum, there seems to be some evidence that humans perceive ADM as less fair because it does not comprise all aspects perceived as necessary. The case of algorithmic discrimination was not raised, while the problem of algorithmic sensitivity towards errors was. Interestingly, while ADM is praised for its capability of procedural fairness and treating everyone the same, this was not necessarily what most experimental subjects perceived as fair. Instead, there seems to be a preference for exceptions to the general rule, even though humans agree that this would constitute a deviation from the principle of procedural fairness treating everyone the same. An absence of concerns that machines could be discriminatory was also noted in previous qualitative work (Woodruff et al., 2018).

In finding a fairness metric that best matches human notions of fairness, a perceived less fairness of ADM could affect the ratings because humans distrust algorithms in general. So, based on our findings thus far, the answer to the question of whether ADM is regarded to be more fair by experimental participants than human decision-making is best summarized informally by "Yes, ADM is fair according to what you mean by fairness, but it is not really what I mean by fairness". Hellberger and Araujo put this as "fairness is not justice" referring to many other aspects that humans find just just (Helberger et al., 2020).

But the broader concept of fairness as understood by humans that emerged from the review also benefits from the perspective of IS and management researchers. The wish for the availability of human intervention fits into the picture of what is known in IS and management science from the literature on algorithm aversion (Dietvorst et al., 2016). Human intervention on ADM by single actors was also documented in the public sector, what they denominated as upstreaming done by "street-level bureaucrats" (Veale et al., 2018). To this related is the issue of employees'

fairness perceptions in hiring (which was not included in our initial literature search). An ethnographic study accompanying a rollout of an AI hiring system at a large company and documented many examples where human interventions on the "neutral" algorithm were requested: lowering the threshold for the previous intern, letting applicants pass that were just on par off with the critical threshold, or allowing for different thresholds across countries because application numbers were different (van den Broek et al., 2019). All these interventions result in an unequal process because the threshold was different for different instances and did occur through human intervention and not the algorithm. So, while the availability of human control over algorithms might increase the adoption of algorithmic decision-making, there also might be a risk for manipulation by single agents from what they perceive as fair. This problem is also related to "fairwashing" of machine learning models due to their intransparency (Aivodji et al., 2019).

Another topic is the role of ML and software developers in ensuring fairness Cowgill et al. (2020).

To summarize, the seemingly innocent question of "Human or AI" involves many future research opportunities, such as developing a conceptualization of fairness and system design.

1.6 Findings from our Review

Humans have very complicated perceptions about what constitutes fairness in a particular situation. Moreover, these perceptions include considerations that are not covered by the fairness metrics.

Fairness perception can be improved if the possibility of human intervention or overwriting of ADM is included in the process. There are, however, risks of (involuntarily) manipulation of the ADM through human actors.

1.6.1 Implications for Research

We list the implications for future research:

- Current fairness metrics may not be exhaustive.
- Scenarios are sensitive to many factors. Therefore, there may be a need for a scenario bank containing calibrated and parametrized situations (e.g., similarly as the

information systems community already uses the Inter-Nomological Network for identifying construct identity (Larsen & Bong, 2016).

- There is a need for a better conceptualization of fairness preferences in algorithmic decision-making.
- Enhancing fairness also means thinking about the costs of wrong decisions carefully. Our taxonomy of gateway and selection decisions may be helpful.
- There are few studies on fairness perceptions in algorithmic hiring, even though this is a topic of interest for the broader IS community (van den Broek et al., 2019).

1.6.2 Conclusion

Recently, the concept of algorithmic fairness has gained traction. However, what is the most preferred metric of fairness? A few studies have been undertaken to crowdsource fairness perceptions to determine the statistic with the highest association with layperson fairness perceptions. We summarized the current literature on that topic. We aimed to provide an overview that other researchers might utilize to perform similar crowdsourcing experiments. For this, we also reviewed the experimental procedures because, to the best of our knowledge, as the topic is relatively new, not so much is known yet about how to do research intersecting machine learning and human perceptions.

Additionally, we explored some of the drawbacks to such undertakings. That is, we discussed the ethical implications of crowdsourcing fairness perceptions. Here, it is essential to address the target population to ensure algorithmic fairness. Furthermore, we also discussed the circumstances and possible dangers of human intervention in ethical machine learning.

2. Human Perceptions of Fairness: A Survey Experiment

Sengewald, Julian; Schlichter, Anissa; Siepermann, Markus; and Lackes, Richard, “Human perceptions of fairness: a survey experiment” (2023). Wirtschaftsinformatik 2023 Proceedings. 72.

3. AI-assisted Learning Feedback: Should Gen-AI Feedback Be Restricted to Improve Learning Success? A Pilot Study in a SQL Lecture

Julian Sengewald (TU Dortmund)

Mathis Wilz (TU Dortmund)

Richard Lackes (TU Dortmund)

Abstract. Generative AI has great potential for ubiquitous access to individualized instruction, which has important ramifications for equal access to education. However, educational theories suggest that constant guidance can impair deep thinking skills. We aimed to evaluate how to best provide AI support, hypothesizing that faded guidance facilitates better transfer learning. In a pre-registered experiment, students were randomly assigned to learn SQL with an AI tutor, providing either full-access to assistance or access that faded. The results were compared on tasks assessing memorization and transfer skills. Contrary to the hypotheses, students supported by the non-restricted AI tutor demonstrated significantly stronger learning gains on memorization tasks. These findings offer novel insights into the role that generative AI plays in optimizing learning outcomes and how to implement it into the learning process.

3.1 Introduction

The trend towards digitalization has had a strong influence on education, with the increasing use of digital tools that complement traditional teaching methods. This movement is fueled by the decrease in expenses associated with digital education, along with the emergence of new technologies that lead to the development of new digital learning tools. Digital learning tools offer students continuous opportunities for practice and formative feedback. Unlike traditional classroom assessments that focus on skill examination, formative feedback aims at nurturing skill development (Gedye, 2010). A digital tool often used in practice are AI tutors (Hobert, 2019). These tools encompass a variety of advantages that make them a promising digital learning tool. An AI tutor reduces the workload of human teaching assistants because it provides automated feedback to student learners. They provide instant, flexible feedback that is easily accessible, enhancing learning efficiency and accessibility. Students can request feedback from an AI coding tutor at any time of the day, allowing them to schedule study sessions at a time that fits more easily into their daily schedule. Additionally, they offer a comfortable and non-judgmental environment for students to ask questions and receive feedback. Also, AI coding tutors can give students feedback on their questions, i.e., when they

are too shy to ask a human teacher. Overall, AI tutors seem to have the potential to enhance educational quality. In addition, these systems also represent a chance to increase educational equity because of their relatively lower costs compared to human instructors. A traditional form of digital tutoring system is an intelligent tutoring system. Intelligent tutoring systems are based on branches of AI such as knowledge-based systems, NPL, and classification systems (Weber et al., 2021). Intelligent tutoring systems are designed by experienced instructors and have a canned set of knowledge and feedback messages (Mitrovic, 2012; Wollny et al., 2021). A new form of technology is generative AI (GenAI). GenAI systems have the capability of producing new content and answering questions. In contrast to traditional intelligent tutoring systems, GenAI has a richer knowledge base and is also able to generate a variety of feedback messages. These tools can provide extensive support to students without the pedagogical expertise of human instructors. The question is whether this development is good or bad for the development of students' critical thinking. Although technology in many ways has improved human flourishing and freed humans from doing unpleasant or harmful tasks, technological progress has also led to concerns about the deskilling of human expertise (Eliot, 2021; Lu, 2016; Sambasivan & Veeraraghavan, 2022). The effect of deskilling, a reduction of human expertise in coping with certain complex problems, together with the problem of over-reliance on AI tools (Buçinca et al., 2021), highlights the importance of balancing AI assistance with opportunities for learners to develop their problem-solving skills. This raises concerns that students who rely too much on AI could misjudge their learning progress by overestimating it (Prather et al., 2023). Therefore, the balance between opportunities and the deskilling effect is crucial in educational contexts, as it could reduce the development of the metacognitive skills of the students. To improve our understanding of the benefits and dangers of GenAI for human expertise, we study GenAI in education in a field experiment in an educational context, particularly focusing on student learning and critical thinking. Furthermore, we provide data on the effectiveness of our proposed instructional design that facilitates the integration of GenAI into digital learning. For this, we report the findings of an experimental study that emerged from a critical stance on the use of generative AI in intelligent tutoring systems. Research on human-AI interaction suggests the concept of cognitive forcing strategies enhancing collaborative decision-making (Buçinca et al., 2021; Gajos & Mamykina, 2022). Building on this, our research investigates how these strategies can improve learning. In contrast to the literature on AI-human-decision-making, this work focuses on cognitive mechanisms to improve human learning when accompanied by AI tutors. This work aims to experimentally evaluate via a field experiment whether limited access to generative AI feedback affects the

learning success of memorization and transfer tasks. Contrary to our assumptions, we found no significant effects between participants in transfer tasks. Participants with unrestricted access performed better on the memorization tasks. Overall, we were able to show that unrestricted access to an AI tutor leads to higher usage and interaction frequency. Our work therefore provides additional empirical data on the integration of generative AI in classrooms to support student learning. In addition, it adds nuance to discussions about the balance between assistance and independence in technology-enhanced learning environments.

3.2 Background

3.2.1 How Humans Learn

The literature encompasses various theories from different disciplines that explore the relationship between instructional practice and its impact on student learning.

The Cognitive Load Theory (CLT) posits that the working memory of learners is limited, and therefore learners can only deal with a certain number of cognitive processes simultaneously. Cognitive processes are experienced as cognitive load. According to the CLT, cognitive processes take part in three channels: intrinsic cognitive load, extraneous cognitive load, and germane. *The intrinsic cognitive load* is related to the complexity of the inherent task. Inherent task complexity is determined by element interactivity, where element interactivity determines the independence of different elements to be learned. If elements can be learned independently (e.g., vocabulary), the intrinsic complexity is lower, whereas if the elements are highly dependent, the task complexity is high. *Extraneous cognitive load* is linked to instructional design. Poor instructional design increases the extraneous cognitive load, so fewer resources are left for the learner to deal with the intrinsic cognitive load. *Germane cognitive load* refers to the extent to which a learner can shift cognitive resources to cope with the intrinsic load. Therefore, the *Germane cognitive load* is experienced as an active process by the learners in contrast to the passively experienced intrinsic and extraneous load (Klepsch & Seufert, 2021). Germane cognitive load is linked to learning (Renkl et al., 2004). More experienced learners may perceive the complexity of a task as less challenging than novice learners because they have become more familiar with the concepts to be learned and therefore the element interactivity is lower for these learners (Sweller, 2010).

Overall, CLT theory differentiates between the cognitive load that can be manipulated by the instructor (intrinsic/extraneous) and the germane load, which fosters learning. In pedagogy, *scaffolding* refers to the temporary support provided by an instructor that enables a learner to

accomplish a task they could not complete independently. The *Zone of Proximal Development* (ZPD) theory explains the progressive enhancement of learners' problem-solving capabilities. Problems are categorized into three zones: those solvable independently, those insoluble by the learner, and an intermediary zone known as the *zone of proximal development*, which necessitates assistance. With appropriate support, the learner's problem-solving capacity can be progressively expanded by instructors. Instructors provide support that is *contingent* on learners' needs. Learners with little prior knowledge are given more support than more advanced learners, whereas extensive support is less beneficial for advanced learners. A further common characteristic of scaffolding shared across the literature is the concept of *fading* (Belland et al., 2017, p. 319; van de Pol et al., 2010, pp. 274–276). Fading refers to the practice in which the extent of support is decreased over time and responsibility is gradually transferred to the learner. Fading can also be provided only at students' discretion (self-selection) or at fixed times (Belland et al., 2017).

3.2.2 Feedback in Learning

According to prior literature, there exist several theoretical models of feedback in education (Lipnevich & Panadero, 2021). For its broad applicability, the Hattie-Timperley model (HT) is often chosen to encompass a variety of instructional practices (Lipnevich & Panadero, 2021). According to the HT model, the content of feedback can take four different forms: task focus (how well the task was executed), process focus (how the task can be completed), self-regulatory focus (addresses commitment, control, and confidence towards the learning goal), and self-focus (Hattie & Timperley, 2007). The HT model thus specifies the content of instructional-effective feedback.

Empirical studies present divergent findings regarding the most effective modality for delivering feedback. Delaying the *timing* of feedback, for example, appears to lead to better test performance, although it is unpopular among students and is believed to degrade test performance despite the empirical results hinting otherwise (Mullet et al., 2014). *Elaborative feedback* refers to the question of how detailed feedback should be, contrasting it conceptually from instruction. Providing elaborative feedback during a learning phase in an explanatory style (explaining why something is wrong) yields higher test performance than benchmark groups that only received correct/wrong answer feedback. This effect can be seen both in repeated and new exam questions, although in the latter case, the difference in effect is greater for elaborative feedback (Butler et al., 2013). Another form of feedback, often encountered in digital learning environments, is *comparative feedback,* which assesses how a learner performs compared to

their peers (Günther, 2021). Another form is *gamified feedback*, which tries to entertain learners, such as using appealing avatars to motivate them (Schneider et al., 2018). Gamification in learning can have positive effects on learner's progress, but it can also increase the cognitive burden and extraneous load for learners, leaving less cognitive resources for productive learning (Cai et al., 2022). However, the extraneous load imposed by overly gamified interfaces can also outweigh the distraction incurred by adding irrelevant aesthetic content if this content is enjoyable and motivational (Skulmowski & Xu, 2022). In summary, research on educational interventions provides some examples where teaching approaches intended to benefit students proved less effective than more conservative methods. The divergent results across feedback modalities and settings illustrate the complexity of determining optimal feedback delivery.

3.2.3 Intelligent Tutoring Systems

While human instructors typically rely on teaching strategies that they are either explicitly or implicitly aware of, it is more difficult to build algorithms that exhibit the same human expertise. Constraint-based tutors were one of the first attempts to achieve that objective. By representing domain knowledge as constraint rules, these systems analyze students' solutions to identify errors and provide feedback on violated principles (Mitrovic, 2012). Intelligent tutors have been used in a variety of forms, content-wise and modality-wise (Mitrovic, 2012). They have been used for programming assistance or mathematics education (Mitrovic, 2012). Recent technological advancements have made it easier for non-programmers to create content or use classification/mining algorithms for developing intelligent tutoring systems (Wambsganss et al., 2021; Winkler et al., 2020). For all these approaches, however, the development of actual content is an essential step to be performed (e.g., dialog flows, and labeled training data).

Recently, much research has been undertaken on the form of presenting intelligent tutoring systems. Intelligent tutoring systems are increasingly deployed in the form of conversational agents (Hobert, 2019; Hobert & Wolff, 2019; Wollny et al., 2021, p. 8). Conversational agents can take the form of a chat-like interface or an embodied entity supporting the learner's progress (Hobert & Wolff, 2019). It is noteworthy that chatbots and conversational agents have been a widespread form of delivering pedagogical support in roles such as assistive, educative, or mentoring (Wollny et al., 2021) long before GenAI arrived in the popular discussion. Another form of delivery is multi-modal delivery. *Multi-modal* delivery uses, besides text-based interfaces, also audio for educational assistants to improve learning because learning can be

distributed on several cognitive channels simultaneously (Winkler et al., 2020). Other forms of delivery include dynamic scaffolding intervention via accessible smart personal assistant systems which proved more effective in improving problem-solving skills compared to students who had no access to these tools (Winkler et al., 2021).

3.2.4 Generative AI and Generative AI in Education

Generative AI (GenAI) models represent a relatively new branch of AI capable of producing new content. These models take the form $p(x|c)$ where $p(.)$ refers to a probability distribution on some data x conditional on covariates c. Within the AI literature, they are described as *conditional generative models* because the output depends on the input provided. Applied GenAI usage consists of developing suitable inputs to achieve the desired output, referred to as *prompt engineering*. However, GenAI requires significant resources for training. Typically, publicly available pre-trained machine models like Chat-GPT (Open AI), Bard, or Llama are used in practice. These models are built for general purposes but can often be readily used for special domains with minimal adjustments. GenAI's versatility lies in handling diverse input data and imitating structures, making it ideal for complex text generation and programming tasks.

Unsurprisingly, GenAI also is used in education with broad implications. One of the prominent concerns about GenAI is the potential for academic dishonesty, where students plagiarize content with the help of GenAI. While the academic content creation capabilities of GenAI exist (Kasneci et al., 2023; Megahed et al., 2023), some have argued that GenAI may be more viewed as a supplemental to academic writing (Bishop, 2023). Students are certainly aware of GenAI as a supportive tool (Prather et al., 2023). On the perspective of educators, the usage cases for Chat-GPT in education range from helping educators to prepare course materials, to the creation of exam tasks (Lo, 2023) or grading (Dai et al., 2023). GenAI like ChatGPT-3 can provide feedback with consistently high readability, although performing better at process-oriented oriented feedback than task-oriented feedback (Dai et al., 2023). The performance of ChatGPT-3.5 in programming tasks depends highly on the programming language (Megahed et al., 2023). Overall, we did not find a study on the capabilities of GenAI to give feedback to students in a real educational setting and its influence on learning outcomes important in higher education.

3.2.5 Summary of Contribution

Compared to prior research on intelligent tutoring systems (Hobert & Wolff, 2019; Khosrawi-Rad et al., 2022; Wollny et al., 2021), the novelty of this research simultaneously lies in: (1)

the *underlying technology* used for the AI tutor; (2) the *design construct*, a cooldown timer for the availability of feedback, which should nudge students to think deeply on their own, while not risking them getting stuck and hindering learning progress by giving appropriate feedback (Cai et al., 2022; Gedye, 2010; van de Pol et al., 2015); (3) The *scope of the study* is the study of the effect of AI feedback modes on two different outcome measures. These outcome measures encompass the assessment of reproductive knowledge (memorization) and the assessment of metacognitive skills (knowledge transfer), which are important in higher education, especially in advanced-level undergraduate courses compared to introductory courses.

3.3 Hypothesis

Learning new subjects is always challenging for students. If the new subject has a relatively high conceptual distance from the prior knowledge, making cognitive links between the new material and the existing knowledge becomes more difficult. *Element interactivity* (Sweller, 2010) is a theoretical concept that educators can use to understand how learning can be improved, especially when teaching new and complicated topics. In complicated subjects, such as teaching programming, element interactivity is high because learners are unfamiliar with the material and perceive the new concepts as novel and distinct from their current knowledge. Typically, in programming, there is simultaneously freedom and convention with respect to how to put a conceptual problem into appropriate computer code. However, novice learners often struggle to distinguish between these elements during code creation (e.g., order of execution, naming, syntax). Additionally, in programming the concepts interact with each other. Errors are often symptomatic of more complex cause-and-effect relationships. The novelty and complexity of programming topics lead to high-element interactivity.

CLT states that high-element interactivity leads to a high cognitive load for learners. Instructors can reduce the extraneous cognitive load for students so that sufficient resources are left for learning to occur. One leverage point based on instructional theory presents *worked examples* (Barbieri et al., 2021; Renkl et al., 2004). In line with CLT and scaffolding, providing worked examples and support should benefit students learning because it frees up cognitive load for students to focus more on the learning task at hand.

When measuring the learning outcome, we tested the student's ability to solve SQL tasks. To do this, we distinguish between memorization and transfer tasks. By tasks that require transfer skills, we mean content that is not necessarily new to the learner but is performed in a different environment or under different conditions, which can be considered as the 'application'-

dimension of Blooms Taxonomy (Krathwohl, 2002). In the SQL context, for example, this refers to exercises that are carried out on a new database and require the transfer of previously learned commands to a new database schema. Transfer tasks can also occur when SQL concepts are used in a combination that students have not encountered during their practice time. Transfer tasks thus require more deep understanding, because they require more deep thinking activities. In contrast, we understand memorization skills as the ability to solve tasks that have already been performed in the same or a similar way. In this case, memorization tasks can be associated with the 'remember'-dimension of Blooms Taxonomy, as they have been completed in exactly the same or similar way by the students and therefore only need to be recalled (Krathwohl, 2002).

To investigate whether AI feedback impairs learning effects in memorization tasks, we divided the study participants into two groups. The first group has full-access to the AI feedback, while the second group has a time limit before new feedback can be requested. Here we use the concept of fading in the context of scaffolding. The restricted group is referred to as fixed-faded feedback by self-selection.

***H1**: For memorization tasks that do not require deep understanding, there will be no significant difference in learning outcomes of a non-faded (full-access) and a faded AI tutor (partial-access).*

On the other hand, providing too much support that is not *contingent* on the learner's needs may be harmful to the learner's success (Renkl et al., 2004). The *reversal of worked example effects* in line with CLT suggests that advanced learners, who can solve more difficult problems requiring deep understanding, need less support because they already possess sufficient cognitive schemes about what the study (Kalyuga & Renkl, 2010) In this case too much feedback can lead to reversal effects by blocking necessary mental resources for idea generation. The *theory of skill acquisition* suggests that integrating practice, which turns declarative knowledge into procedural knowledge, with worked examples that necessitate cognitive active processes on behalf of the learner by self-explanation than no self-explanation improves learning outcomes (Chi et al., 1989). Cognitive psychology research suggests that learning is optimal when there is a certain level of resistance implemented, increasing active cognitive engagement, by establishing a *desirable difficulty* level and employing *self-testing* and *interleaved practice* which enhances both retention and comprehension of the subject (Biwer et al., 2020; Brown et al., 2014). In the context of learning with an AI tutor, learners could tend to outsource tasks, which are considered too difficult, although these would be

exactly the right level of difficulty to achieve learning progress. Also, knowledge is often a byproduct of practice, referred to as *incidental learning.* Incidental learning is the unconscious acquisition of knowledge during practice (Gajos & Mamykina, 2022). Experimental research shows that learners who self-direct their decisions, a *cognitive forcing function*, based on feedback and explanations, benefit more from incidental learning as measured in subsequent tests (Gajos & Mamykina, 2022). This suggests that cognitive engagement during the learning phase leads to improved retention and application of knowledge. Therefore, learners that were suitably cognitively engaged through an appropriate cognitive forcing function, in our case the interruption of feedback necessitating an active cognitive process on behalf of the students, perform better when exposed to experimental conditions that facilitate such cognitive engagement due to the *compound effect* of enhanced incidental learning (Gajos & Mamykina, 2022), reduced cognitive load (Kalyuga & Renkl, 2010), self-explanation (Chi et al., 1989) and retrieval-testing (Biwer et al., 2020; Brown et al., 2014):

H2: *For transfer tasks that require deep understanding, learning outcomes will be larger when supported by a faded AI tutor (partial-access) compared to a non-faded AI tutor (full-access).*

To evaluate these hypotheses, we conducted an experiment. The experiment design, hypotheses, and analysis plan were pre-registered on the Open Science Framework prior to data collection[4].

3.4 Experiment

To test the hypotheses, this experimental study used a design with two experimental groups to evaluate the effect of AI-assisted coding help on student learning outcomes. For the experimental groups, we manipulated the access to the AI feedback tool on a self-developed SQL Learning Platform that is used in university-level database management courses.

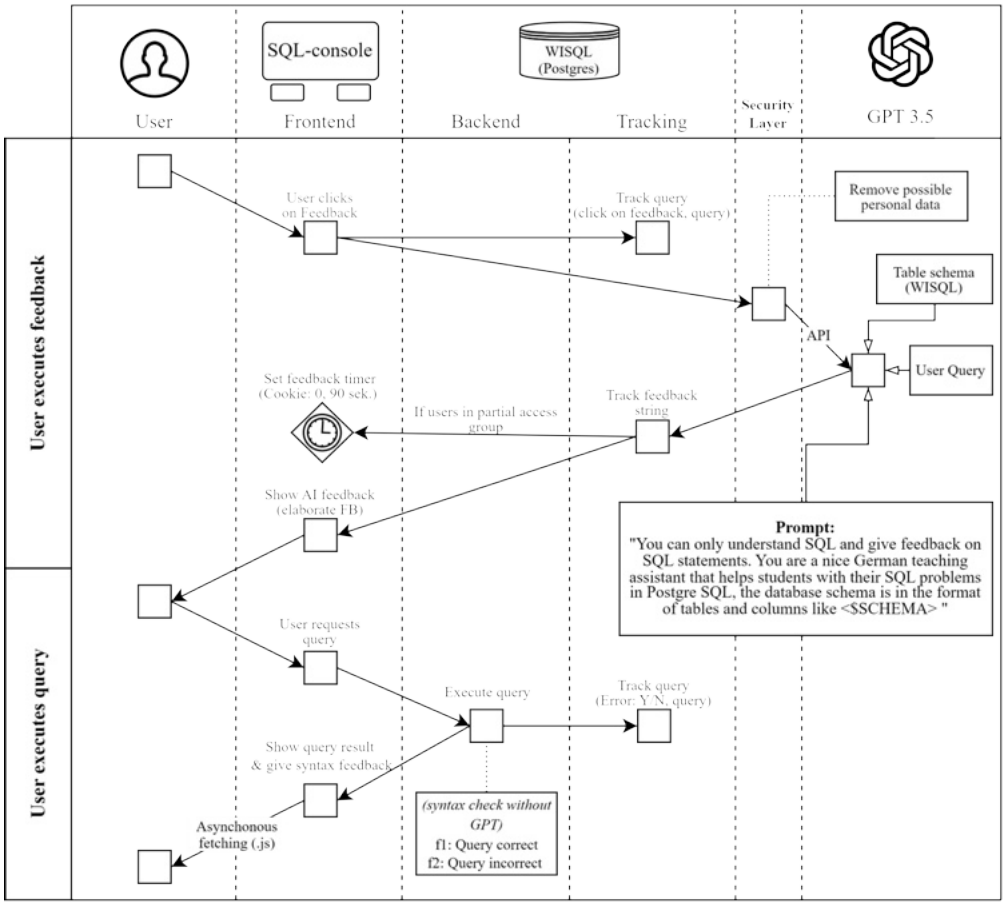
3.4.1 Online Platform

Within this learning platform, students have access to an AI feedback tool that provides suggestions on how to improve their SQL query and avoid errors. The architecture of the platform is depicted in Figure 3.1. Feedback can be requested via a button (element C) and is asynchronously fetched into the page as a displayed text message (element D). The GenAI feedback tool is powered by OpenAI's GPT-3.5-turbo API. The SQL query, schema, and instruction prompt are passed through to the API. With this information, GPT can give feedback

[4] https://osf.io/ab2g9/?view_only=cf52455e4ba2411291b22b6a934e5469

to the students to query related questions. The use of the feedback tool is optional. The platform consists of a frontend in which students find an interface that shows them the SQL database schema. A console input area is connected to a PostgreSQL database in the backend. The platform is a modernized version of a learning platform developed at the chair of one of the co-authors. For this experiment, we built an AI coding tutor that has components in the backend and in the frontend (Figure 3.2). The frontend utilized JavaScript offering AI feedback depending on which experiment group students belong to. The backend contained (1) a security mechanism to prevent disclosure of personal identifiable information and (2) functional logic to provide GenAI-assisted feedback. When students submitted queries to the AI tutor by clicking the feedback button in the frontend, the feedback was generated by passing an instructive prompt to GPT-3.5-turbo enriched with contextual information about the backend. This prompt contained relevant context like the student's original query, the database schema they were working with, and any error messages or status outputs from PostgreSQL. By structuring the prompt this way with diagnostic information, the generative model was able to produce adaptive, explanatory feedback tailored to each learner's specific query or issue.

Figure 3.1: Platform architecture

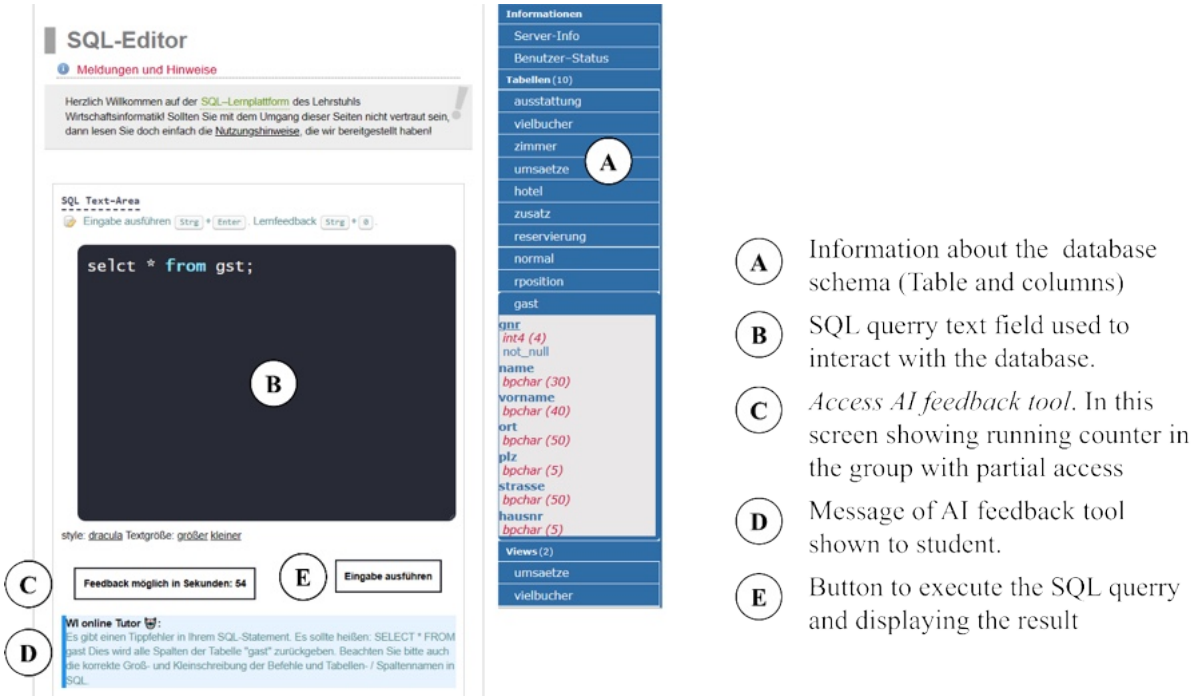


The feedback was asynchronously loaded into the fronted by JavaScript so that there was no need to reload the page which could have interrupted the students' learning experience.

The AI coding tutor is used in two conditions to give elaborative feedback. These conditions affect the timing of availability of support provided by the AI coding tutor tool during students' studies. The two levels are:

1. **Partial-access**: Every time a student demands AI feedback through interacting with the interface he/she must wait 90 seconds before receiving AI feedback again.
2. **Full-access**: Students have no time limit on AI feedback and can request assistance as often as they wish.

Figure 3.2: SQL Learning Platform



In this case, the study follows a scaffolding approach that allows participants to self-select the moment of feedback, which is fixed-faded for 90 Seconds for the experimental group. The size of the interval corresponds on average to 20% of the planned completion time per task. In this case, the timer was intended to prevent students from trial and error.

3.4.1 Procedure

Using the above setup, the experiment was carried out over several weeks accompanying an in-presence undergraduate database management course and involved activities on and outside the platform. An overview of the experimental plan is given in Figure 3.3.

The experimental groups were formed according to standard protocols (Dennis & Valacich, 2001, pp. 16–21). The students were randomly assigned to one of the two experimental factor

levels (stratified sampling). The assignment considers the gender and SQL experience (self-reported) of the students to ensure an even distribution over both groups.

For three weeks, two groups of students were assigned identical SQL homework tasks using an online SQL Learning Platform equipped with an interface, database access, and an AI feedback tool. At the end of the three weeks, an assessment of students' SQL skills was conducted. The assessment consisted of a 30-minute pen-and-paper exam that was designed to evaluate memorization and transfer of learning through various SQL tasks. During the exam, students did not have access to the AI feedback tool or a computer. Both groups were tested simultaneously at the same location. The assessment exam encompassed familiar database content covered through the exercises, as well as a genuinely new database schema to assess students' capability to cope with a new problem. To be eligible to take part in the exam, students had to dedicate at least four hours to practicing on the SQL Learning Platform. This ensured that all participants were adequately exposed to the experimental conditions. During the practice period, students received two SQL homework sheets containing a series of exercises revolving around the same database. The sheets did not contain procedural solutions (i.e., SQL syntax); for more difficult exercises the desired returned tables were included as a form of self-control. The second sheet was sent out about a week after the first and contained more advanced exercises. The first sheet was given to the students before the start of the study and before registration.

The students were encouraged to participate in the study by offering bonus points, which could improve their final exam score, as determined by their achievement in the 30-minute SQL exam. The graders of the SQL test were unaware of which experimental group the students belonged to, and the students were unaware that they were participating in an AI feedback study. The score on the 30-minute SQL exam will be used as the principal outcome measure.

We assume that the incentive offers a medium-strong incentive. Students can receive up to 6 additional points on the final exam, counting 90 points depending on their performance in the test exam. The bonus points provide an incentive to complete the final test and comply with the experiment. By this approach, we can ensure that students put effort into their performance and preparation for the test exam. This helps the exam results to reflect the student's true abilities of the student more accurately.

Additional control variables collected during the two-week study period include the intensity of tool usage, the number of SQL syntax errors, the total study/prep time, and the number of

times feedback was requested. These measures are collected by the SQL Learning Environment.

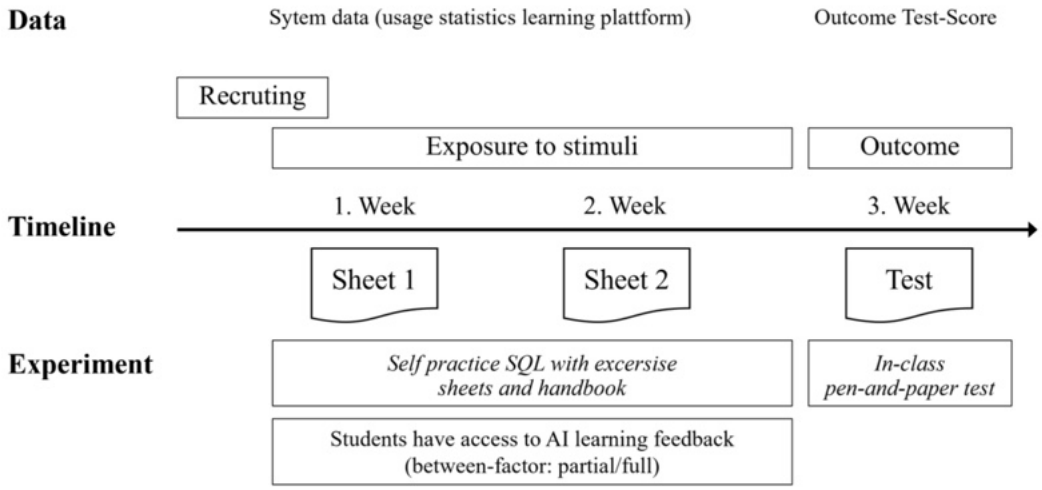


Figure 3.3: Experiment Timeline

Both exercise sheets used for training had a total of 32 tasks. All tasks were, in our opinion, easy to moderately difficult. We measured the difficulty in the complexity of the queries. This means that the more nested a query is (e.g., number of joins), the more difficult it is. We would call queries that combine multiple joins with an additional query as 'difficult'. In total, it was possible to complete the exercises in around four hours. In addition, the students were given a handbook that was tailored to the exercises and explained the SQL-basics and some information about the user interface and the login process. The final test was about the same level of difficulty and included single-choice and free-text tasks.

3.5 Empirical Results

In this study, we conducted multiple statistical tests to investigate potential differences between the two groups, using a multiple-dependent variable method to ensure reliable results. We start with the overall results and subsequently delve into more specific outcomes.

3.5.1 Sample

The sample consisted of all students who took part in the mock exam. The attrition of the participants who signed up for the experiment until participation in the experiment´s mock exam corresponds to 38%. Leaving a sample of 21 students (10 in partial, 11 in full). The gender ratio is highly skewed with respect to gender with most of the study participants being male. The course is targeted at business students but is also open to computer science and business mathematics students. Most participants in the mock exam stemmed from the latter two subjects.

3.5.2 Main Outcomes

3.5.2.1 Results

The participants in the full-access condition achieved higher scores in the mock exam ($M = 27.455, SD = 4.156$) in comparison to those assigned to the partial-access condition ($M = 24$, $SD = 6.7$). The Shapiro-Wilk test result ($W = 0.919$, $p = 0.063$) suggests that the data´s distribution is not significantly different from a normal at conventional ($a = 0.05$). The Welch-t-test used to compare the mean scores did not indicate a significant difference ($t(14.771) = 1.403, p = 0.181$, ns), suggesting that the variable controlling access to AI coding help does not have a significant impact on students. To further assess these findings, we performed a non-parametric Wilcox test. The Wilcox test did not show statistically significant differences between the conditions ($Z = 70, p = 0.304$). These findings were consistent with the t-test. Both parametric and non-parametric analyses indicated that there was no discernible distinction between the two groups in terms of their overall test scores.

We also hypothesized that there is a higher effect with respect to the impact of AI coding help on tasks that require transfer whereas tasks that are just pure memorization should not be affected. For memorization tasks, our hypothesis predicts no significant difference across the two conditions. Participants in full-access performed better on memorization tasks ($M_{\text{full, memorization}} = 18.091, SD_{\text{full, memorization}} = 0.944$, $M_{\text{partial, memorization}} = 15.6, SD_{\text{partial, memorization}} = 3.565$, $CI_{memorization,0.9} = [0.4,\ 4.6]$). The Shapiro-Wilk normality test ($W = 0.755,\ p < 0.01$), indicates that the data significantly deviates from a normal distribution and the parametric test would be more appropriate. The non-parametric Wilcoxon test ($Z = 73.5,\ p = 0.188$) yielded a non-significant result. For reference, we also report the Welch-t-test comparing the mean scores between the two conditions did reveal a significant difference ($t(10.146) = 2.142, p = 0.057$). Parametric and non-parametric analysis suggest different results concerning the difference between the two groups for memorization tasks.

Contrary to the second hypothesis, that students in the partial condition would perform lower on transfer tasks compared to students in the full condition, our findings in the descriptive statistics indicate otherwise ($M_{\text{full, transfer}} = 7.545, SD_{\text{full, transfer}} = 2.841$, $M_{\text{partial, transfer}} = 6.5, SD_{\text{partial, transfer}} = 3.171$, $CI_{transfer,\ 0.9} = [-1.2,\ 3.3]$). The Shapiro-Wilk normality test ($W = 0.89,\ p = 0.0158$) indicates that the assumption of normality for the Welch-t-test is fulfilled. The Welch-t-test comparing the mean scores between the two conditions did not reveal a significant difference ($t(18.204) = 0.793, p = 0.438$). The non-parametric test ($Z =$

$68.5, p = 0.35$) yielded a similar result. Both parametric and non-parametric analyses suggest that there is no observable difference between the two groups.

3.5.2.2 Discussion

Contrary to expectations, the full-access group performed better than the partial-access group in the test. Study conditions (full and partial-access) were assigned randomly. Nonetheless, the small number of participants in the study suggests that the results should be interpreted with caution. Therefore, we have reported confidence intervals for the group difference estimate. The confidence interval for the difference in memorization task $CI_{memorization,0.9} = [0.4,\ 4.6]$ indicates all positive values, suggesting a genuine effect. On the other hand, the confidence interval for the difference in transfer task $CI_{transfer,\ 0.9} = [-1.2,\ 3.3]$ also includes zero and negative values, suggesting there may be no difference between the groups or that the difference between the groups may be even negative.

3.5.3 Control Variables

3.5.3.1 Results

Before obtaining the mock and final exam scores, we collected extensive usage data that encompassed participants' engagement with the learning platform for both groups. Participants in the full-access condition ($M_{\text{full}} = 28.5, SD_{\text{full}} = 24.928$) used the feedback function, measured through the number of clicks, less often than the partial-access group ($M_{\text{partial}} = 31.167, SD_{\text{partial}} = 19.385$). With respect to time in hours spent on the platform, the effect was reversed ($M_{\text{full}} = 3.988,\ SD_{full} = 1.611,\ M_{partial} = 3.20,\ SD_{partial} = 1.48$). The statistical tests showed no significant difference between the two groups for both variables ($t_{\text{feedback}}(df_{\text{feedback}}) = 11.957, p_{\text{feedback}} = 0.826; t_{\text{time}}(df) = 18.995, p_{\text{time}} = 0.256$).

We also analyzed the student traces of the log data. For this, the table compares inter-event times (in seconds) between console inputs and feedback requests for students with full versus partial-access to GenAI tutor. Students in the full-access condition of GenAI exhibited shorter gaps between console inputs (134s average) and feedback requests (109s average), with smaller standard deviations. In contrast, partial-access elicited longer intervals between interactions (157s for inputs, 193s for feedback) and greater variability per the higher standard deviations. Statistical tests confirm that the between-group differences are highly significant ($p < 0.01$) for both interaction types.

Table 3.1: Usage statistics of the platform and corresponding results in the mock exam.

Inter-event time (in seconds)	**Full-access**		**Partial-access**	
	Average	*Std. Deviation*	*Average*	*Std. Deviation*
Console Input	134***	262	157***	283
Feedback	109***	242	193***	324
Test results	*Average*	*Std. Deviation*	*Average*	*Std. Deviation*
Score	27.46	4.16	24.00	6.70
Memorization tasks	18.09	0.94	15.60	3.57
Transfer tasks	7.55	2.84	6.50	3.17

3.5.3.2 Discussion

For the control variables study time and the number of times the feedback was requested, there were no significant effects. This non-significant finding suggests that the absence of an effect on the primary outcome variable, test score, is unlikely to be attributed to attendant disparities in the control variables. The indirect effects on the results, resulting from the influence of treatment on study behavior, could be a concern for the validity of the results. For instance, the AI feedback tool's engaging appearance and interactive nature could be perceived as enjoyable, motivating students to study for longer durations when they have access to the feedback tool more frequently. In this scenario, the treatment would primarily manipulate study time rather than promote deep thinking. However, since there were no significant differences in total study time between both groups, this suggests that indirect influences are not the primary driving force behind the observed outcomes.

Following this line of thought, we further investigated whether study time had any effect on the outcome. Study time can also be viewed as the length of exposure to the treatments. For this, we defined post-hoc a low and high-usage group with usage of more or less than 4 hours as a cut-off. The parametric test yielded a significant difference for the users with low usage ($t_{\text{low}}(df_{\text{low}}) = 6.727, p_{\text{low}} = 0.076$), but not for the high-usage ($t_{\text{high}}(df_{\text{high}}) = 6.156, p_{\text{high}} =$

0.638). The non-parametric test was not significant ($W_{\text{low}} = 11, p_{\text{low}} = 0.634, W_{\text{high}} = 21, p_{\text{high}} = 0.138$). The p-values were adjusted for multiple comparisons using Hochberg-procedure.

In general, all participants were required to have some exposure to treatment, as the incentive (bonus points) was conditioned on the usage of the platform (at least 4 hours of practice time or finishing all tasks). The results suggest that study time had a differential effect depending on the amount of usage. Study participants with low exposure to the treatment showed an effect contrary to our expectations, as the full-access group performed significantly better than the partial-access group. In the high-usage group, this effect is reversed. This reversal may be an artifact due to an outlier and small sample size in the high-usage group rather than an actual effect of high exposure to the treatment.

The time lag between each request for feedback (inter-event time feedback) showed that the partial-access group took approximately 90 seconds longer on average between feedback requests than the full-access group, confirming the intended manipulation. This observation provides validation for the experimental control, as there was a discernible distinction in inter-event timing. Additionally, the full group's average inter-event time of 100 seconds indicates the delay of 90 seconds for the partial-access group was not set too low since we doubled the time that passed before the group requested feedback again leaving them twice the time compared to the immediate condition to engage in deep thinking. This means the experimental manipulation to engage students in deep thinking worked as intended.

The observed group difference (Table 3.1) also suggests that the interval was not too restrictive because if it had been too long, both conditions would have performed similarly regardless of the assigned condition. Therefore, our experimental manipulation induced, in principle, a well-calibrated waiting interval to foster comparatively deeper cognitive engagement for the partial versus full-access group.

The standard deviation for inter-event time (the time between feedback requests) was around 30% larger in the partial-access group than in the full-access group. Larger standard deviations suggest more variability in the time students waited before requesting feedback under the partial-access condition. A possible reason is that students in the partial-access spent more time thinking on their own for exercises of different difficulty levels, even after the feedback timer expired, especially for more difficult tasks that needed more thought. On the contrary, students in the full-access groups appeared to request feedback at more regular intervals, implying more consistent help-seeking behavior.

3.6 General Discussion, Limitations, and Take-Aways

This study explores how digitalization enhances classroom learning. We evaluate GenAI as a coding tutor and test it in a real setting. Previous research work used coding tutors that were not based on GenAI and were built specifically for their intended purpose (Hobert, 2019; Wollny et al., 2021). GenAI, in contrast, facilitates instructional support to students with minimal effort required. In our specific case, we were able to implement the system in just a few days. To achieve this, we passed the students' SQL query, along with the SQL schema, and a deliberately constructed prompt, to a GenAI API. This allowed us to create a coding tutor that offers flexible feedback to students.

Previous work, e.g. (Mitrovic, 2012), used constraint-based rules to build an intelligent tutoring system that gave students feedback adaptive to the situation. According to our understanding of prior research, much work is needed to build intelligent tutors (Hobert, 2019; Khosrawi-Rad et al., 2022). GenAI is a new available technology before inaccessible to educators. By applying GenAI in our proposed way, one can significantly reduce the effort needed for personalized instruction. Our results show that GenAI tutors are effective for students.

There also can be formulated a critical stand on GenAI besides the worry of plagiarism. We, therefore, evaluated our system concerning different learning outcomes and explored the best way to provide GenAI-assisted feedback based on instructional theory. In particular, we examine the implications of providing intelligent AI-enabled learning support to students and delivering this support in a contingent manner where the responsibility for learning success is gradually transferred to the learner (van de Pol et al., 2015). For developing a pedagogical effective tutoring system there is the need to balance support without diminishing students' critical thinking skills (van de Pol et al., 2015). As generative AI tools have not been vetted by experienced instructors, they risk providing excessive assistance that fails to cultivate independent problem-solving. Given the general availability of generative AI tools to learners, this poses the risk that students lose important metacognitive skills such as critical thinking and independent problem-solving. So far, the use of GenAI use in education has undergone only limited testing in classrooms. Previous work (much outside IS) evaluated the performance of GenAI in educational applications (e.g. replacement of human teachers) (Dai et al., 2023; Lo, 2023), but paid scant attention to critically reviewing its drawbacks or empirically evaluating remedies.

Our study adopts a dual perspective: recognizing GenAI potential while scrutinizing challenges like the problem of over-assisting students. To address this, we designed a low-invasive self-

selected fixed-faded scaffolding intervention for the generative tutor that nudged students toward critical thinking through regulated feedback, without imposing overly restrictive limitations. The goal was to nudge these student groups into deep thinking on their own rather than just passively consuming feedback from the GenAI tutor. Controlling feedback timing is aimed to shift responsibility for sense-making temporarily to students through productive struggle and self-explanation, which are important for developing problem-solving and transfer skills (Chi et al., 1989; Schworm & Renkl, 2007).

To encourage self-explanation, we used regulated feedback through a 90-second timer in one of the experimental groups (partial-access). The timer imposed periodic delays before learners could access additional feedback from the AI tutor. Surprisingly, we found that the group that had full-access to the feedback tool performed equally or even better for the transfer tasks than the control group. This finding implies a discrepancy between the experimental results and the pertinent theory. We did not find that forcing students to think deeply on their own about the possible source of mistakes led to increased problem-solving capability in a delayed examination of the skills they had acquired.

Previous experimental research on digitized scaffolding has demonstrated its effectiveness in enhancing student learning. Specifically, scaffolding, when compared to non-scaffolding-based instruction, has been shown beneficial for both memorization and transfer tasks (Winkler et al., 2020). It also has been associated with problem-solving skills (Winkler et al., 2021). As this previous research, we also studied the effect of scaffolding on memorization and transfer tasks. However, our focus was on comparing two variants of providing scaffolding (restricted vs unrestricted GenAI support) drawing insights from established educational theories (Chi et al., 1989; Schworm & Renkl, 2007; van de Pol et al., 2015).

Our findings implicate plenty of support leads to higher educational success than restricting it, at least under the conditions of the current experiment. That means AI coding tools can support learners also in advanced undergraduate courses. How our results extend to the graduate level (e.g., master programs) that require more abstract thinking and self-regulated learning from students would remain an open issue. It could be that AI coding tutors may be not beneficial in these cases, which has been observed as the expertise reversal effect in other contexts (Kalyuga & Renkl, 2010). One potential explanation for our results is that the students lacked sufficient mastery and less from deep individual thinking. This means the assignments' element interactivity remained high, explaining in line with theory (CLT and scaffolding) why extensive support still helped more than in the corresponding control group.

A source of interference for the findings could involve the design of feedback timing. In this study, we set it to 90 seconds. Such an interval may have been too short, considering that students need time to read and process feedback. Thus, the nominal restrictive time interval is the result of the difference between 90 seconds and the student's cognitive processing time. If the processing time is high, there would be no difference between the two experimental groups, even though the timer was used (our empirical results suggest that the timer was not too high). In addition, the timer should also not be set too short, because the processing of feedback represents mental effort and is influenced by factors such as the complexity of the information, the individual's familiarity with the topic, and their cognitive abilities (Sweller, 2010). The complexity of the information was the same for all students. The GenAI tool could not adapt to how often students had requested feedback as it had no memory of past interactions. Furthermore, given that the course is at an advanced undergraduate level, most learners are familiar with the SQL syntax. Therefore, the learners had some prior knowledge, and the programming content of the feedback messages should not significantly increase their cognitive processing time. A higher interval would also not represent a practical viable option. If intervals are too long, students may passively wait for the timer to expire rather than think deeply, browsing elsewhere, and prolong the time before re-engaging with learning. Overall, this could decrease learning. Thus, a restrictive limit may not prove pedagogically effective. Alternatively, feedback could depend on demonstrated student effort, e.g., requiring worked solutions first. However, readily available GenAI (Prather et al., 2023) could allow circumventing behavior. A feedback design that is too restrictive may push learners outside of the learning platform. Therefore, optimally promoting learning while not over-restricting poses a clear inevitable trade-off for faded coding tutors in real education environments.

As with every study, this study has limitations that should not be unmentioned. First, most participants in the sample were male. In this case, there were eight out of eleven male participants in the full-access group and nine of ten in the partial-access group. In general, more males take the elective advanced-level courses we used for the experiment. Yet, the gender ratio more strongly shifted towards male students for reasons that we cannot explain. On the downside, our results mostly extend to male students, but it also leaves a more homogeneous sample. Another limitation is self-selection in the sample. First, participation was voluntary, and we cannot observe how the effect would have been for sample outsiders. A similar problem is attrition, which could be attributed to motivation and subject mastery. Students who felt after some weeks in the course that they would not perform successfully in the graded exam also have no incentive to take part in the mock exam. Finally, the present study only looked at

programming skills, especially SQL. Other studies look more at AI tutors, yet not GenAI-based, specifically for developing argumentative writing skills (Wambsganss et al., 2021). The effect of faded tutoring as implemented by us may differ from other educational domains. Finally, the study is restricted to the scope of the sample itself, which is based on an advanced-level undergraduate course at a German university. Effects at the introductory undergraduate or graduate level may differ.

After the general discussion we will describe the main results of the study are as follows:

- Conceptualized a low-cost approach to building AI coding tutors by using off-the-shelf generative models rather than full custom development.
- Empirically evaluated an AI tutor integrated into an authentic undergraduate classroom, providing ecological validity for the results.
- No learning gains from timed feedback regulation for problem-solving, contrary to the hypothesis of a trade-off between assistance and autonomy.
- Full-access to the AI tutor increases usage, interactivity of usage, and frequency of practice.

3.7 Conclusion

While there is an emergent debate in education about the question of whether to ban GenAI, this debate should be informed by data and good theory. This study set out to take a critical stand on this debate, considering potential gains and losses. Given that education is about preparing students for a future that is still unknown in its exact contours but likely to differ from the present, educators should do their best to equip students with the skills and knowledge necessary to thrive in a changing world. It is likely that students, in their future labor, will collaborate with GenAI in ways that we cannot imagine yet. Hence, if this future includes GenAI, there is no good reason to forbid its use in education. For this study instead, we conjectured that GenAI will affect students' skills in a way that might harm them eventually more than it benefits current learning progress because overly relying on GenAI feedback could reduce their metacognitive skills, especially critical and problem-solving skills. If educators seek to restrict GenAI usage, this restriction should be justifiable. The fact that GenAI reduces work for the current student generation ('that is unfair') is not a well-founded reason. Looking back in history, search engines, including academic databases, have taken away much of the work of pouring over library catalogs that before was considered an essential part of academic effort, yet nobody would deny the merits search engines have had. Much more, effective

education cultivates skills that have value beyond the classroom. A particular problem likely to become relevant in the future is the problem of AI over-reliance. Overly using GenAI may thus not only be a problem through deskilling once students take part in the workforce (Eliot, 2021; Sambasivan & Veeraraghavan, 2022) but manifest itself much earlier. If left unaddressed, over-reliance on AI suggestions risks creating learners unable or unwilling to independently reason through problems themselves. One compelling reason to restrict GenAI is to address the issue of over-reliance on artificial intelligence. By adequately teaching students to independently perform tasks, they can better identify errors when working alongside AI. While the problem of excessive reliance on AI is acknowledged within the IS community (Buçinca et al., 2021), it seems that the potential complexities of using GenAI for providing intelligent feedback in education have not been thoroughly examined yet.

Contrary to our hypothesis based on educational theory (and personal intuition), our data suggest that we should let students use GenAI at their will because it improves learning outcomes, including problem-solving skills. Letting students explore learning at their own pace is a particular instance of practice-based learning. A particular problem with this learning approach is that students can get stuck, hindering their learning, if left completely on their own. GenAI-assisted learning tutors provide a way for students to receive instructional support at their discretion. Using its low-cost, GenAI can support learners, resulting in more equitable access to education due to AI tutors facilitating the realization of an enhanced, continuously available system of learning guidance. This has important implications since many students who are working besides their studies attend face-to-face classes less often and thus have less chance to receive feedback from human instructors. Furthermore, this study suggests that there are benefits associated with access to GenAI learning assistants and the templates set up in this study are likely to be beneficial for further research and teaching practice.

II. Understanding Inequity in Machine Learning Systems: A Data-Centric Approach

While a thorough understanding of fairness in ML systems is crucial from both a theoretical and behavioral perspective, it is equally important to examine the specific mechanisms that can lead to unfairness in these systems. This chapter focuses on the role of data as a key input in ML systems and explores how it can contribute to unfair outcomes. The current chapter explores data based sources of unfairness as they arise in complex ML systems. In Figure II.1, the conceptual differences between the two contributions are illustrated.

The first article in this section (C3) takes a sociotechnical perspective, examining how user behavior can lead to unfair ML systems and how operators can mitigate these effects. It looks at the role of data at the record level, analyzing how the removal of individual data points (e.g., through privacy legislation) can perpetuate bias and lead to discriminatory outcomes.

The second article in this section (C4) takes a technical perspective, focusing on the interplay between privacy and fairness at the microdata level. It explores how data sharing can facilitate better insights while raising concerns about privacy and individual rights. It examines how these concerns can be addressed to ensure responsible and ethical use of data in ML systems.

By examining data at both the data set and aggregate levels, this chapter provides a comprehensive understanding of how data can contribute to injustice in ML systems. It offers valuable insights for researchers, practitioners, and policymakers working to develop and implement fair and equitable ML solutions.

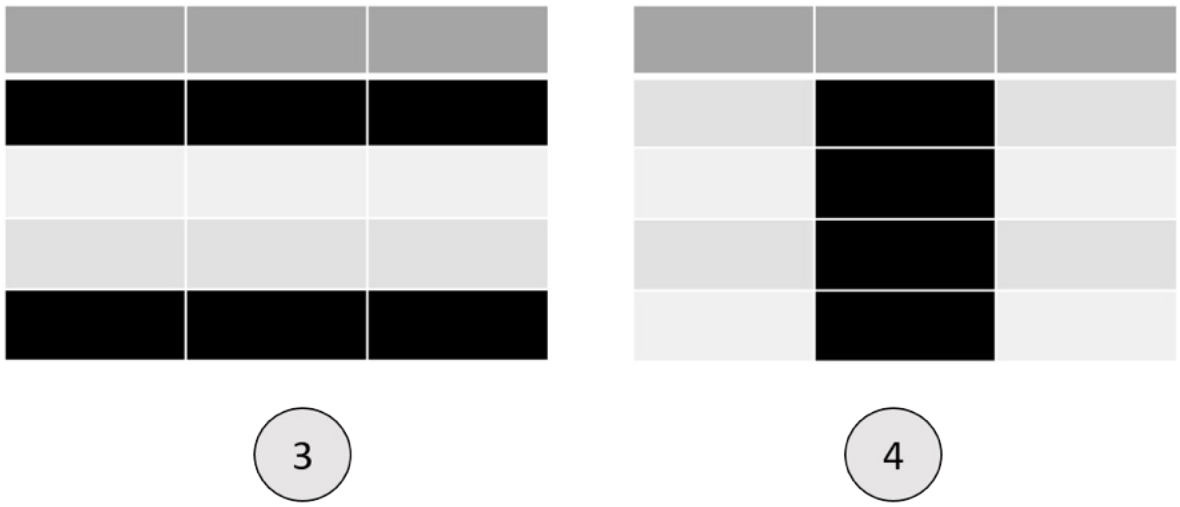


Figure II.1: Comparison of conceptual differences between contributions

4. The Impact of the 'Right to be Forgotten' on Algorithmic Fairness

Julian Sengewald (TU Dortmund)
Richard Lackes (TU Dortmund)

Abstract. Enterprises may often deal with situations in which they cannot use a part of their data for machine learning: (1) Privacy laws grant users to determine that their data should not be used any longer by the data holder (i.e., a data record must be removed). (2) Privacy laws may require anonymization, which leads to the removal and masking of specific attributes from a dataset. (3) To avoid ethical issues, it may be necessary to prevent machine learning algorithms from using attributes in their decision-making that may be viewed as discriminatory (i.e., a feature may not be used because it is unethical). Given these challenges, we analyze how removing data from an information system in the cases mentioned above affects its predictive performance and the fairness of its decisions.

4.1 Introduction

Personalized recommendations based on machine learning is a widely used technology in digital services (for example, displaying personalized news or providing users' favorite music based on their preferences). At the same time, many digital services are built on the assumption that their users will provide data to fuel the machine learning engine, resulting in more accurate personalized recommendation capabilities. However, public opinion on the legal regulation of machine learning systems is changing. Numerous laws governing data gathering and machine learning technology have been enacted, affecting systems that are already in use. For example, the EU General Data Protection Regulation (GDPR) in 2016, the UK Data Protection Act 2018, India's Personal Data Protection Bill in 2018, and the California Privacy Rights Act 2020. This places digital service providers in the challenging position of having to engineer a technical solution to meet the legislation's standards for privacy and personal data, but also to ensure that the personalized recommendations are still useful.

The implications of these laws are potentially not just limited to privacy, because machine learning depends on the data that the individual users provide. Some regulations allow self-selection (tracking, privacy consent) and self-redaction of personal data (e.g., personal data deletion request). In general, self-redaction occurs when consumers request that their data be deleted from a digital service and self-selection occurs when the user chooses how much data

is collected. Users may choose to share selectively just certain types of information (Dinev & Hart, 2006; Phelps et al., 2000) or to remove (Milne & Rohm, 2000) (e.g., because of growing concerns about data breaches or increased privacy awareness). Selective data sharing and deletion may also occur if consumers experience discrimination based on the data they provide and prefer instead not to share such information.

In practice, groups of users may only disclose a portion of their digitally available information to a service provider. Some examples are: A person may desire to keep their gender hidden to avoid gender discrimination. Similarly, their ethnic origin (Gunarathne et al., 2019). In all of the aforementioned cases, the user will either restrict data sharing or ask the digital service to erase their personal data. For data analysts, all of these possibilities culminate in the fact that they only deal with a subset of available data and that certain data have been manipulated or anonymized.

But, what are the consequences of limited data availability on machine learning model-based decisions in digital services under selective data sharing and different technical solutions for implementing the 'right to be forgotten'? Two factors must be considered: the overall quality of the analytic results (e.g., classification performance), on the one hand, and how well the machine learning systems perform for each demographic group, on the other.

Two factors must be considered: the overall quality of the analytic results (e.g., classification performance), on the one hand, and how well the machine learning systems perform for each demographic group, on the other. The second factor is frequently referred to as algorithmic fairness (Feldman et al., 2015; Hardt et al., 2016). We hypothesize that minority groups suffer from limited data availability because the remaining group members cannot fully compensate for a loss of training data, and hence unfairness may arise. Our contribution emphasizes the impact on the fairness of the predicted outcomes. Fairness is an increasingly important concern in developing automated systems and algorithmic decision-making, especially in recent years (Barocas & Selbst, 2016; Haas, 2019; Latanya Sweeney, 2013). The paper analyzes the empirical evidence from a simulation study on machine learning automated credit lending decisions to address the research question:

RQ: *What is the effect of different technical solutions for implementing the 'right to be forgotten' on algorithmic fairness for minority groups and overall predictive capability in machine learning based decision-making?*

The conceptual background on legal requirements for data deletion and anonymization is reviewed in Section 4.2.1 - Section 4.2.4, and the cybersecurity problem in machine learning

models is reviewed in Section 4.2.2. We then proceed to non-discrimination in automated systems and the special category of data according to GDPR in Section 4.2.4. and the measurement of fairness in Section 4.2.5. Related literature is summarized in Section 4.2.6. In Section 4.3 we develop our hypotheses and present the technical details of our analysis in Section 4.4.1. The Section 4.5 contains the empirical analysis. The paper concludes with a discussion of its results Section 4.6.

4.2 Conceptual Background

4.2.1 The right to be forgotten

Two provisions included in the GDPR grant users the right to control their data once transferred to a digital service. First, the well known 'right to be forgotten' is founded on Article 17's user's withdrawal of consent for lawful data processing and entails the complete removal of their data. Second, Article 18 gives data subjects the right to restrict the processing of their personal data, but does not entail data removal. However, if the data is restricted from processing, it can not be used for machine learning any longer and hence has the same result as removing it. However, strictly deleting all records from the database can cause additional technical burdens Villaronga et al. (2018). Furthermore, as data deletion processes must comply with a wide range of different regulations and laws to be lawful in all jurisdictions, the complete deletion of a record may be unlawful as enterprises may also be obliged to preserve certain information (e.g., the GDPR itself states some exceptions when data must not be deleted). Anonymization is a reasonable choice for such a circumstance because the impacted records will not be deleted entirely from the database. Instead, all the data about the person is masked in a way that the data record can not be linked back to the actual individual. The enterprise may pursue different degrees of data masking to be compliant with different legal requirements or regulations. Furthermore, if the data is masked to be considered anonymous, then data deletion requests can be rejected (Austrian Data Protection Authority, 2018). The GDPR requires data controllers to give their subjects the power to select where and how their data is handled. On the other hand, data controllers can reconcile privacy rights with other organizational goals using anonymization techniques.

Similar regulations to those found in the GDPR can also be found in other laws Table 4.1.

Table 4.1: Legislation on data protection in different countries.

Name	Legislation	Link
Indian Data Protection Bill	Article §25 and §27	https://www.meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf
California Consumer Privacy Act	Nr. 1798.105	https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.105#
UK Data Protection Act	§46-48	https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

4.2.2 Anonymization

Anonymization removes all the personally identifiable information from a database to prevent identification. The identification risk assessment should consider both current and potential technological developments that will increase the risk of identification. Therefore, enterprises may opt for a higher degree of anonymization to also reduce the risk of identification in the future.

Proper anonymization requires at least the removal of any direct identifiers from the data records. Since direct identifiers are often discarded because they usually do not carry any information that can be used for machine learning, therefore deleting these identifiers does not amount to any information loss for machine learning. However, a subset of attributes may require different handling. Quasi-identifiers (QIDs) are attributes where a single QID may not be sufficient to identify a person, but several QIDs together might be used to identify all or at least some records. In such circumstances, the privacy model of k-anonymity can be employed to suppress or generalize some QIDs as needed to maintain privacy protection (L. Sweeney, 2002a, 2002b). Because QIDs are often used in machine learning models, their distortion may harm machine learning performance.

4.2.3 Cybersecurity

The cybersecurity of public machine learning models is complicated by the fact that they are typically trained with personal data. The so-called model inversion attacks can expose specific properties of training data (e.g., certain features), whereas membership derivation attacks can reveal whether an individual is a member of the training data (Fredrikson et al., 2015; Shokri et al., 2017; Veale et al., 2018). Furthermore, the security of machine learning services

accessible via an API is vulnerable (Tramèr et al., 2016). Because of this, some researchers doubt a machine learning model's anonymity (Veale et al., 2018).

Owing to these security concerns and the relationship to privacy laws, enterprises are confronted with the risk implied by storing sensitive data in a machine learning model. As a result, some data may be removed from publicly accessible machine learning models to prevent leakage of sensitive personal data.

4.2.4 Non-discrimination and special categories of data

Algorithmic fairness seeks to prevent people from being mistreated in automated decision-making based on their membership in a particular group. These groups include, for example, those defined by age, gender, and interests (Speicher et al., 2018; Latanya Sweeney, 2013). In practice, many deployed machine learning models have been found to perform substantially differently across various demographic groups (Köchling & Wehner, 2020; Latanya Sweeney, 2013). Therefore, businesses may want to avoid utilizing particular groups in machine learning models for ethical concerns. Also, the GDPR and other laws agree that certain sensitive data require specific handling (technical or processual). For example, some categories of sensitive data are not allowed to be processed (e.g., Art. 9 No. 10 GDPR). Similar data types are defined in other legislative systems (Electronics, 2018; United Kingdom, 2018).

Because of the aforementioned regulatory restrictions, organizations may remove sensitive data from their machine learning programs to comply with privacy or non-discrimination regulations.

4.2.5 Fairness measurement

Technical fairness metrics quantify the extent of discriminatory treatment of a particular group. These metrics measure how much weight machine learning systems put on protected attributes. Protected attributes are characteristics of a person that should not be used as the basis for decisions as determined by law, courts, other authorities, or moral reasons. The protected attributes can signify membership to particular groups, as stated in Section 4.2. 4. The domain of the protected attribute is then the set of all possibly existent groups.

For simplification, we assume that there are only two groups: g_1 is the group for which a possible unfair treatment is of concern relative to the treatment that group g_2 receives. The prediction of the machine learning system is denoted by $h(x)$ for any individual/record x. If $h(x) = 1$ the machine learning system will allocate a benefit based on the information contained in the record x. For example, $P(h(x) = 1 | \, x \in g_1)$ is then the probability of how

often the machine learning system will allocate a benefit to the members of the group g_1. We define g_1 as the possibly disadvantaged group. Using this setup, the following fairness metrics can be defined:

The demographic parity metric. The demographic parity (DP) metric considers an allocation unfair if it disproportionally allocates a benefit to one community:

$$DP = \frac{P(h(x) = 1 \mid x \in g_1)}{P(h(x) = 1 \mid x \in g_2)}$$

The criterion relates to the disparate impact criterion, also known as the 4/5 rule, in the US labor law (Feldman et al., 2015). Concretely, if $DP < 0.8$ then a disparate impact (DI) occurs for which a company can be held legally accountable if the allocation can not be justified.

It is possible to broaden the measure's scope to include whether or not the benefit recipient is eligible. For example, in supervised machine learning, this knowledge would be carried by the real class label y(x) where $y(x) = 0$ are non-eligible, and $y(x) = 1$ are eligible for any record x.

Equalized odds (ED). A fairness metric that conditions the actual outcome predicted by an automated system is the equalized odds constraint (Hardt et al., 2016). Let the true-positive-rate $TPR(g_1) = P(h(x) = 1 \mid x \in g_1, y(x) = 1)$ and the False-positive-rate $FPR(g_1) = P(h(x) = 1 \mid x \in g_1, y(x) = 0)$. Then equalized odds requires:

$$TPR(g_1) = TPR(g_2) \quad \text{and} \quad FPR(g_1) = FPR(g_2)$$

Equal opportunity (EO). Only requiring equal predictive outcome for the positive class leads to the fairness criterion equal opportunity (Hardt et al., 2016):

$$TPR(g_1) = TPR(g_2)$$

Equal opportunity is a less stringent fairness constraint than equalized odds. Only positive recall must be the same across the groups.

Although DIs lends itself to a directly quantifiable metric, equal opportunity and equalized odds are defined as a fairness constraint. To quantify deviation from the optimal fairness situation in the latter case, we define unfairness as:

$$Unfairness^{+}(g_1, g_2) := \left| 1 - \frac{TPR(g_1)}{TPR(g_2)} \right|$$

For the negative recalls, $Unfairness^{-}(g_1, g_2)$ can be specified accordingly. Thus, one obtains a metric for measuring how equal opportunity and equalized odds fairness constraints are

violated. Using the absolute value in the preceding equation results in the favoring and disfavoring of the group $g_1$over g_2 are weighted in the same way.

4.2.6 Related literature

In the context of the right-to-be-forgotten, the term 'machine unlearning' emerged first in a setting where one had to develop an efficient technique to erase data from a huge machine learning system because retraining would have been too time-consuming (Köchling & Wehner, 2020). Further work investigated the viability of design principles for developing efficient deletion algorithms and considered an unsupervised learning problem (Ginart et al., 2019). Using these principles, the authors created two k-means algorithms for efficient data deletion. Other work was directed towards providing guarantees for linear classifiers that ensure that the machine learning model behaves as if it had never been exposed to the data (C. Guo et al., 2020).

Aside from record deletion, some work has been done on removing sensitive features from a machine learning model. In such a setting, robust submodular maximization methods were used to efficiently eliminate features from a model after being trained since it is not always known before model training that a feature is sensitive (Kazemi et al., 2018).

Precautionary de-personalization (under the restrictions of privacy laws like the GDPR) is another approach to handling the right-to-be-forgotten. There is a wealth of literature dealing with disclosure control techniques for data publishing (e.g., Chamikara et al., 2020; Domingo-Ferrer & Torra, 2001; Fung et al., 2010; Karr et al., 2006; Mivule & Turner, 2013; Soria-Comas & Domingo-Ferrer, 2018).

Finally, some work has been done on implementing the right-to-be-forgotten for complex data structures, such as blockchains or streaming/distributed data (Farshid et al., 2019; Kazemi et al., 2018).

Despite having a wide variety of available technologies, little study has been done to examine how data deletion affects other societal objectives. This is unfortunate as privacy laws may be designed to protect human rights and interests. One of the few studies of unintended consequences of enforcing privacy in machine learning is Chen et al. (2020). They find attacks executed before and after data erasure give an adversary insight into the erased record's sensitive information. Thus, data deleted from the machine learning model is eventually disclosed, resulting in an outcome entirely contrary to what was intended.

4.3 Hypothesis Development

Enterprises may face different situations in which they need to delete a part of their data from machine learning to comply with privacy regulations. We refer to these strategies as personal data protection strategies (PDDR) strategies. By enumeration, we give four strategies a digital service may pursue on a technical level:

1. **Record deletion.** Deleting the complete record, which may be the most straightforward strategy.
2. **Record masking.** Only masking, anonymizing, the record(s) for which a data removal was requested (compliant with previous jurisdiction Austrian data protection authority 2018). Masking can be seen as a soft deletion in that it erases certain details while leaving other details intact.
3. **Anonymizing the complete table.** Performing proactive anonymization of all records so that the corresponding regulation of data removal does not apply.
4. **Deleting a single sensitive attribute or collection of sensitive attributes.** For example, if such sensitive features are typically often requested to be removed, it may be practical not to use them in the machine learning system at all so that the model does not need to be retrained. Furthermore, the omission of such attributes may also be considered ethical or may mitigate privacy leakage if a cyberattack comprises the machine learning model.

The consequences of PDDR strategies on algorithmic fairness of data deletion in the context of a selective-random PDDR mechanism have gone mainly unexplored. By the selective-random PDDR mechanism, we refer to the case when data deletion requests are overrepresented within a particular demographic group.

The literature investigated different reasons why machine learning models may discriminate. For example, the data being used for machine learning is imperfect by itself, and the machine learning model may learn to reproduce already existing discrimination in the machine learning model (Barocas & Selbst, 2016). In addition, some features may be less reliable or more inaccurately collected for a part of the population or just not a good predictor for some (minority) groups (Barocas & Selbst, 2016; Mullainathan & Obermeyer, 2017). Also, feedback loops can degrade the fairness of machine learning. A feedback loop occurs when the machine learning model decisions also affect what training data the machine learning model can see in

the future, which leads to 'skewed training examples' (Barocas & Selbst, 2016; Ensign et al., 2018). Based on these considerations, we developed our first research hypothesis:

H1: *Record deletion and masking lead to increased unfairness.*

When data is deleted, the machine learning model has fewer examples to learn from. If a minority group was already more difficult to learn, since a minority group encompasses by definition fewer instances, eliminating data only exacerbates the issue, increasing existing errors. If more errors occur only in one group, the TPR will decrease for this group and increase overall unfairness. Similar arguments can be made for masking, which reduces the amount of information in a specific value for an attribute. However, we expect the loss of information caused by masking a single QID or a collection of QID to be less than the loss of information caused by removing records:

H2: *Record deletion causes higher unfairness than record masking.*

We model two separate scenarios, low and high, to differentiate further the impact of different data deletion mechanisms on fairness. In scenario low, only a small number of deletion requests come from a potentially disadvantaged group (e.g., the machine learning system is actually discriminating or wrongly perceived to be discriminating, and users, therefore, request for the deletion of their data). In scenario high, a larger number of data removal requests stem from the potentially disadvantaged group.

If people, who are eligible or qualified for a particular benefit allocated by a machine learning model (e.g., university admission, loan application, or job offers), are discriminated against or perceive the possibility of their discrimination because of affiliation with a group, they could request data removal. The ability to request data removal reduces the number of cases the machine learning model can see certain characteristics linked to a particular outcome, leading to more unsatisfactory overall performance for that group. We suppose that data deletion induces a special case of 'skewed' training examples (Barocas & Selbst, 2016). We thus formulate the following hypothesis:

H3: *The greater the number of users requesting personal data be removed, the worse the impact on fairness and predictive performance.*

Anonymizing the entire table has an approximate equal impact on all groups and makes it more difficult for the model to learn from them. That is, for example, the group-specific recall would decline for all groups and thus would not affect primarily a single group-specific recall. However, if there is any discontinuity between group-specific sample size and group-specific

error, the recall for a marginalized group may decrease further, which is only partially offset by the comparatively more minor increase in-group-specific error (i.e., lower recall) for the other group. In that case, the unfairness may increase slightly:

H4: *Anonymizing the whole table affects unfairness only to a small extent.*

Deleting a feature from a table may be closest to the notion of fairness that usually involves human decision-making (e.g., many countries have privacy protection and anti-discriminatory laws prohibiting companies from requesting photos of job seekers).

However, while this technique may help to minimize unfairness in cases where the machine learning model could use a particular attribute to discriminate against a marginalized group (i.e., the inability to reproduce discrimination that is already existing in the world through the distribution of the label attribute (Barocas & Selbst, 2016), the effect only takes place if that information can be removed entirely from the dataset. In many cases, this may not be feasible because data is usually multivariate, in the sense that attributes are not only associated with the label attribute but also with each other. That is the case with redundant information. Consequently, the machine learning model may deduce the excluded feature from the remaining features in the dataset (Shokri et al., 2017).

H5: *Removing a feature from the table may reduce unfairness if no redundant information is contained in the database, and, quite seldom, it will increase it.*

4.4 Simulation Study

The study was designed to test the hypothesis that deleting a subset of data records, values of a set of attributes of a single record, or entire attributes would lead to a loss of fairness (PDDR strategies). For this, we conducted a simulation study.

4.4.1 Implementation of PDDR strategies

In particular, record and feature deletion is just directly implemented by erasure from the dataset. Masking is implemented as follows: Categorical attributes may be masked with a dummy value such as "*", and numerical attributes are imputed (as if they were missing data) using a naïve mean imputation procedure. We randomly include some "*" already in the training data to avoid model retraining. Anonymizing the whole table was carried out by the k-anonymity model on the QID (i.e., generalizing the QID as needed to achieve k-anonymity), for which we used the datafly algorithm L. Sweeney (2002a). Datafly chooses the QID heuristically for generalization.

4.4.2 Simulated scenarios

We utilized stratified 70% subsampling to generate the bootstrap replicates. The same bootstrap replicates are used to compare each of the three deletion techniques. We also repeatedly evaluate different users that could have solicited a request for data deletion on the same replicates. The number of bootstrap resamples is 500, and the number of different sets of users that request data deletion is three, yielding a total of 1500 evaluations. In addition, the performance of the machine learning model is evaluated using ten times repeated five-fold cross-validation. The selective-random PDDR mechanism is implemented as follows: The probability for a user group soliciting a PDDR is $P(\text{solicit PDDR} \mid x \in g_1) = \tau$ and $P(\text{solicit PDDR} \mid x \in g_2) = 0$. We simulated two different parameter values for τ describing the percentage of users who request data deletion: a low scenario (5% of group members solicited data deletion) and a high scenario (20% of group members solicited data deletion).

4.4.3 Technical details, data, and software used

We rely on the caret package for R to build a machine learning pipeline (i.e., upsampling, cross-validation, scaling, tuning) (Kuhn, 2015.) We use the German credit dataset from the UCI repository as it is a common benchmark dataset for studies in privacy and fairness research. The QID attributes were age, gender, and foreign, for which we developed a suitable generalization hierarchy. Following a setup in previous literature, the fairness-protected attribute was age<=25 years (young vs. old credit applicants) (Haas, 2019).

4.5 Empirical Analysis

4.5.1 Bradley-Terry model

We compared the performance of the machine learning model across different PDDR to address data deletion requests. In particular, we are interested in how predictive performance and unfairness compare across the PDDR strategies and which PDDR strategy is preferable according to those metrics. For this comparison, we used the Bradley-Terry model (Bradley & Terry, 1952.) In the Bradley-Terry model (BTM) β_i denotes the ability of the PDDR strategy i to achieve better performance on a metric γ that an alternative strategy j. The probability that i wins against j is $p_{ij}(\gamma_i > \gamma_j) = \frac{e^{\beta_i}}{e^{\beta_i + e^{\beta_j}}}$. The relative strength of i vs. j is given by $log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_i - \beta_j$. To make the model identifiable one of the strategy coefficients is fixed to zero and all the remaining coefficients are then interpreted as the relative ability compared to the fixated strategy. We fixate when no PDDR strategy is applied. Hence, $\beta_{no\ PDDR} = 0$.

4.5.2 Effects of PDDR strategies on (un)fairness

The results of our simulation study concerning the effect on unfairness are presented in Table 4.2 (note that we do not report p-values separately as due to the large number of simulated outcomes, they are less interesting. However, they are significant on any conventional significance level except for the outcomes marked with "n.s."). Coefficients have to be interpreted in comparison to the case of no handling data deletion requests. The unfairness criterion is $Unf.^{+}(young, old) = \left|1 - \frac{TPR(young)}{TPR(old)}\right|$, and predictive performance is given by accuracy (acc.). We only included simulation runs in the analysis that started with an unfair situation for the young group.

For the DI fairness criterion, which is not reported in the table, only deviations from the threshold of 0.8 are of concern. However, for virtually all methods, this threshold was never violated. Only for table anonymization in 3% of the cases, DI was below this threshold.

Table 4.2: BTM estimates of PDDR abilities β_i (standard error in brackets, $^{ns.}$ p-value >0.1/non-significant, n=1.315).

Variable Setting	**Unf.+ Low**	**Unf.+ High**	$\boldsymbol{TPR^{young}}$ **Low**	$\boldsymbol{TPR^{young}}$ **High**	**Acc. Low**	**Acc. High**
Record Deletion	-0.034 (0.028)	$0.036^{n.s}$ (0.028)	0.259 (0.029)	0.343 (0.030)	0.336 (0.026)	1.191 (0.03)
Record Masking	0.610 (0.029)	0.772 (0.029)	-0.487 (0.030)	-0.637 (0.031)	-0.152 (0.026)	$-0.005^{n.s.}$ (0.027)
Table Anonymization	2.902 (0.047)	2.699 (0.043)	-3.374 (0.061)	-3.551 (0.064)	-1.129 (0.028)	-1.874 (0.037)
Attribute deletion	-1.017 (0.031)	-0.952 (0.030)	1.834 (0.037)	1.624 (0.035)	0.349 (0.026)	0.657 (0.028)

Table 4.3: Overview of hypothesis and results.

Hypothesis	Description	Result	Formula
H1	Record deletion and masking are unfair.	Partially confirmed	$\beta^{Unf+}_{\text{record deletion}} > 0$ $\beta^{Unf+}_{\text{masking}} > 0$
H2	Record deletion is more unfair than masking.	Not confirmed	$\exp\left(\beta^{Unf+}_{\text{record deletion}} - \beta^{Unf+}_{\text{masking}}\right) > 1$
H3	More PDDR leads to more unfairness and worse overall predictive accuracy.	Partially confirmed	$\beta_i^{Unf+,\text{low}} < \beta_i^{Unf+,\text{high}}$ $0 > \beta_i^{Acc,\text{low}} > \beta_i^{Acc,\text{high}}$
H4	Anonymization does not affect unfairness.	Not confirmed	$\beta^{Unf+}_{\text{Anonymization}} \approx 0$
H5	Attribute deletion does not increase unfairness.	Confirmed	$\beta^{Unf+}_{\text{Attribute deletion}} \leq 0$

4.6 Discussion

4.6.1 Results

This research aims to provide light on the subject of which PDDR technique is best suited to ensure fairness. As far as we know, it is one of the first research studies to conceptualize the strategies that a digital service provider might employ to address PDDR. We also discussed the relationship between those strategies and algorithmic fairness and how those two goals might interact. Our findings (see Table 4.2) indicate that there is such a relationship concerning that research questions (an overview of all hypothesis Table 4.3). In particular, H1 is partially confirmed as masking, and but not deletion increases unfairness. H2 is not confirmed as record masking leads to higher unfairness. Our hypothesis that total record loss is always worse than decreasing record informativeness was proven inaccurate here. The masking/imputation procedure impaired the model performance unevenly for the younger group. Therefore, the trade-off between complying with the privacy requirement (and discarding all information about the record) and utility/fairness considerations must be balanced (e.g., utilizing a masking strategy that acknowledges, e.g., the age range).

The comparison between the low and high scenarios shows that most PDDR strategies have higher unfairness in the high setting, confirming H3 except for table anonymization. Contrary to our expectation H4, anonymizing the whole table increased unfairness even though both groups were affected by the k-anonymization procedure. Overall predictive accuracy was also reduced. Differential privacy, a different privacy technique from the one we studied, has also been demonstrated in a different setting to have an adverse effect on fairness (Bagdasaryan et al., 2019; Salas & González-Zelaya, 2020a). Our hypothesis 5 is also confirmed. Indeed, attribute deletion is the method that can reduce unfairness compared to the benchmark case. In so far, attribute deletion turned out to be, in the studied dataset, the best method under consideration. In general, it is also important to consider the target variable's relationship with the attribute being deleted. Younger had a greater default chance in the German credit data set, and, only without, information discriminating between age and predictions for old and young became more similar. As attributes highly correlated with a sensitive attribute may be present in a dataset, attribute deletion is not always recommendable to address fairness (Dwork et al., 2012). Many enterprises might discard demographic attributes entirely from the ML system in an attempt to develop neutral ML systems. This corresponds to our setting of 'attribute deletion' with implications, as pointed out before.

4.6.2 Limitations

In the future, more mechanisms that simulate data deletion requests should be investigated. We only examined the outcome when group-specific data deletion requests occurred. Our methodology is consistent with earlier empirical findings that the distribution of PDDR is sociodemographic disparate (Johnson et al., 2020). Also, we chose a random mechanism $P(\text{solicit PDDR} \mid x \in g_1) = \tau$ and $P(\text{solicit PDDR} \mid x \in g_2) = 0$. A conditional-random mechanism,

$$P(\text{PDDR} = 1 \mid x \in g_1) = \begin{cases} \tau & \text{if } h(x) = 1 \\ 0 & \text{otherwise} \end{cases}$$

, may be possible such that only individuals that did not receive the favorable outcome solicit PDDR. But, many digital services fail to clearly convey to users what the favorable outcome could have been (e.g., displaying a high-profile job advertisement vs. displaying a low-profile job advertisement), and as a result, users may solely suspect if they were discriminated against because they did not see the alternative outcome. Such settings can thus be modeled using our selective-random PDDR mechanism.

The dataset we studied is a well-recognized and well-established dataset in the corresponding research literature (Haas, 2019). Our results may, however, only apply to the specific situation we studied. Nevertheless, we provide a piece of evidence for the design and choice of PDDR strategies.

We only investigated logistic regression. At the same time, machine learning schemes that are transparent and explainable are often preferred over non-transparent machine learning schemes Wang et al. (2020) making a transparent machine learning technique such as logistic regression a sensible default model choice. Therefore, we also focused on logistic regression. However, future research directions may address whether the effects may differ across different machine learning techniques. Finally, we did not study the perception of users towards these PDDR strategies. However, this could be an attractive future research stream.

4.6.3 Implications for practice and research

Several lessons can be drawn from our research. First, different PDDR strategies may lead to different outcomes in fairness and predictive performance, making it important to consider carefully which method may be appropriate. Second, careful implementation of masking procedures deserves some attention as an inferior implementation can also cause unfairness and degrade model performance even though the number of affected records is low. Such a masking procedure should acknowledge the original attribute value of the masked record while remaining anonymous. Finally, practitioners also should consider how PDDR affects the model's performance for the remaining users in the database. In general, such evaluation studies as the ones we described may be used to justify certain PDDR strategies to regulators and national data protection agencies.

5. Balancing Privacy, Fairness, and Utility in Data Sharing: Synthetic vs. Perturbative Approaches a Robust Assessment

Julian Sengewald (TU Dortmund)

Richard Lackes (TU Dortmund)

Abstract. The increase in data collection across domains heightens the need to protect sensitive information while enabling data scientists to build effective machine learning models. This study investigates how organizations can develop responsible data sharing pipelines that balance privacy concerns with data utility for decision-making, and examines the subsequent impact on ML model fairness. To assess the fairness-privacy trade-off across perturbative vs. synthetic privacy-data sharing methods, we conducted a comprehensive computational experiment using standard benchmark datasets and trained 52,800 ML models.

5.1 Introduction

Data sharing is crucial in data science and data governance. Ethical concerns restrict access to data, despite its potential to enhance customer satisfaction, efficiency, and financial value through machine learning (ML) applications. These concerns include privacy rights and *algorithmic bias* (Crawford & Schultz, 2014). Public skepticism about ML's societal implications and heightened consumer privacy awareness also hinder its full utilization. Data governance must address these concerns when sharing data within an organization. These ethical concerns are particularly acute when dealing with sensitive microdata – detailed, record level information about individuals (e.g., customers, employees) in tabular form. Microdata enables powerful applications and digital services, often operated by ML, but it also necessitates robust privacy protections due to its personal nature. Microdata's dissemination poses risks, including objective harms like identity theft and perceived harms from the leakage of sensitive information. Therefore, safeguarding this data is crucial to preventing privacy breaches. Privacy and fairness are fundamental principles for building ethical ML systems (van der Aalst et al., 2017). However, constructing such systems is complex due to the need to reconcile competing objectives. Two significant challenges stand out: the potential for predictive harms from highly accurate ML systems (Crawford & Schultz, 2014; Kosinski et al., 2013), and the tensions between achieving ML fairness and privacy protection (Bagdasaryan et al., 2019b; Cummings et al., 2019; Salas & González-Zelaya, 2020b).

While earlier research has made valuable contributions in understanding the privacy-fairness trade-offs in the ML learning procedure (Bagdasaryan et al., 2019b; Cummings et al., 2019), this research is not directly transferable to *data sharing* contexts where data scientists can freely analyze the data in the form of microdata. Additionally, these studies are mostly theoretical and

do not quantify the typical size of trade-offs (Cummings et al., 2019). Specific research on privacy-preserving data sharing methods suitable for data analysis (Carvalho et al., 2023; Liu et al., 2025) has primarily focused on synthetic data generation methods. Although these earlier works have made significant contributions, they did not fully account for the effects of randomness in data synthetization, dataset composition, and ML training. The first research objective is

RO1: *What systematic patterns of fairness and performance impact emerge when applying privacy-preserving data transformation (PPDT) before ML when evaluated through robust resampling?*

While synthetic data is one significant type of PPDT, perturbative PPDT methods may offer better overall performance in terms of fairness, despite potentially providing less privacy protection (Conde et al., 2024).

RO2: *How does perturbative PPDT compare to synthetic data generation approaches regarding trade-offs between utility, fairness, and privacy for tabular business data?*

5.2 Background

In terms of terminology, a *private dataset* refers to the original dataset without any privacy protection. A *public dataset* undergoes added pre-processing to safeguard privacy, a process known as *privacy enhancement technology*. We call an *adversary* a malign entity with access to the public dataset, but not to the private dataset that seeks to obtain private information about individuals inside the dataset. PET's effectiveness is evaluated by how much private information it preserves and prevents disclosure to the adversary. A *data holder* is a trusted entity (e.g., data governance officer) within the enterprise that is tasked with securely disseminating the data. Let DT be a tabular dataset with N individuals and P attributes. These attributes are ATT $= \{X_1, \dots, X_p, \dots, X_P\}$ and dimension is $N \times P$. The data in this table can be classified into four types: direct identifiers (DI), quasi-identifiers (QID), privacy-protected attributes (PA), and non-privacy-protected attributes (NP). DIs, for example, are full names or IDs that directly identify an individual. DIs are removed. Related to DI are QID. QID differs from DI as only the conjunction of several QID identifies an individual. Examples included age or gender (Fung et al., 2010; L. Sweeney, 2002b). QIDs pose a direct privacy risk as they are observable to an adversary, allowing identification of individuals in the disseminated dataset. The k-anonymity model is used to decrease this risk of disclosing individuals (Sweeney, 2002b). k-anonymity generalizes the QIDs until each unique combination of QIDs defining an

equivalence class, includes at least k records. Hence, identification risk becomes probabilistic and decreases as k increases.

Unlike identity disclosure, attribute disclosure focuses on protecting against revealing private information about specific attributes that an individual wishes to keep private. Attributes for which revealing information is a concern are privacy-protected attributes, PA, and in the reverse case non-privacy-protected attributes NP (Li et al., 2007). The k-anonymity privacy model does not guarantee sufficient protection for attribute disclosure as sensitive information can still be inferred (Li et al., 2007). In this context, (Li et al., 2007) proposed "t-closeness" to quantify privacy protection on PA attributes. Their measure is applicable independent of the measurement scale and incorporates semantic similarity. t-closeness is defined as a measure of privacy on a PA attribute as:

$$t\text{-closeness (attribute}_l) = \max_c \{EMD_{c,l} \mid \forall \text{ equivalence classes } c\}$$

$EMD_{c,l}$ is the earth mover distance (EMD) for equivalence class c on attribute l. EMD is calculated as a linear program finding a flow $f_{i,j}$ between two probability distributions $P_c = \{p_{c,1}, \dots, p_{c,i}, \dots, p_{c,I}\}$, and $Q = \{q_1, \dots, q_j, \dots, q_J\}$, that minimizes the cost of moving an amount of $f_{i,j}$ at costs of distance $d_{i,j}$. The distance $d_{i,j}$ measures the similarity between original and anonymized attribute values.

An alternative framework is the *differential privacy model* (DPM). DPM protects privacy by ensuring that an adversary learns nothing about an individual i by its presence/absence in a query (Dwork et al., 2006). Let $T(.)$ denote a randomized mechanism and D_1 a query containing a specific individual, and D_2 a query without individual i. In a (ϵ, δ)-differential private system $T(.)$ is designed such that the probability of both queries producing a result S is bounded: $\frac{P(T(D_1) \in S)}{P(T(D_2) \in S)} \leq e^{\epsilon} + \delta$. The parameter ϵ controls the level of protection, where lower values imply higher privacy. The additional parameter δ allows for a certain amount of violation of the constraint (Dwork et al., 2006). DPM is a privacy model for synthetic data.

There are two classes of PET techniques, those based (1) on synthetic data generation and (2) those that perturb the original data. Commonly used perturbation techniques involve adding random noise or grouping similar records. In microaggregation, first, a clustering algorithm is applied to the original DT. The cluster size is controlled by the parameter $p_{cluster}$. A clustering algorithm identifies the $p_{cluster}$ records that are the furthest above and below the average to form a cluster and proceed recursively Domingo-Ferrer & Torra (2001). The second parameter of microaggregation replaces attribute-wise the values within each cluster with aggregates for

that group by p_{method} e.g., mean. Noise-based PPDT uses randomness for privacy protection: each record is transformed by an additive noise value. The simplest form of noise adding uses noise values proportionally to the variance of each attribute and a constant factor a (Brand, 2002).

Synthetic data is a PPDT that generates an artificial dataset that preserves the characteristics and information of the original dataset. Generative adversarial neural networks (GANs) are established technologies used to generate synthetic data. GANs consist of two neural networks (NN): a synthesizer and a discriminator. The discriminator classifies data as real or synthetic, while the synthesizer generates synthetic data that cannot be differentiated from real data by the discriminator. DP is enforced on the training algorithm of the NN. Additionally, an aggregation of distributed learned ensembles, known as PateGAN, ensures stronger privacy protection by using several NNs where each is trained only on a part of the data (Jordon et al., 2018). CTGAN, on the other hand, is capable of generating more accurate categorical one-hot-encoded data compared to GANs, utilizing conditional sampling based on the categorical column (Xu et al., 2019).

To compare PPDT s privacy metrics are used (Table 5.1). In addition to being a privacy model, t-closeness is also a privacy metric to compare how effective different anonymization methods are (Li et al., 2007). t-closeness computes how close the PD values in each equivalence class c are to the PA values in the entire data set using EMD as a closeness measure:

$$EMD_{c,l} = \frac{1}{|\text{Equivalence class } c|} \sum_{i \in \text{Equivalence class } c} \left[f\left(x_{i,l;\,\text{public}}\right) - f\left(x_{i,l;\,\text{private}}\right) \right]$$

In this specification, we measure the distance between the distribution of the disseminated public values of the sensitive attribute l, $x_{i,l;\,\text{public}}$, and the global distribution of private values, $x_{i,l;\,\text{private}}$. Here, we suppose that an adversary is interested in learning about the local distribution of individuals within an equivalence class and has access to the global distribution of the sensitive data through some other means (for example, by publicly available census or market research data). The similarity in distribution between the global and local distribution in the disseminated dataset only would not accurately represent actual privacy leakage because all records in the dataset are impacted by a privacy protection technique.

Apart from t-closeness, we use Pearson's correlation as a second comparative measure to quantify the amount of private information, which can be used to quantify the similarity between the original data record and the protected (obfuscated) data record (Kim et al., 2011).

The obfuscated public record is $x_{i;\,\text{public}}$, the original private record $x_{i;\,\text{private}}$ and $\bar{x}_{i;\_}$ the average of its respective attributes.

$$\rho_{\text{private}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left(x_{i;\,\text{private}} - \bar{x}_{\text{private}}\right)\left(x_{i;\,\text{public}} - \bar{x}_{\text{public}}\right)}{\sqrt{\left(x_{i;\,\text{private}} - \bar{x}_{\text{private}}\right)^2}\sqrt{\left(x_{i;\,\text{public}} - \bar{x}_{\text{public}}\right)^2}}$$

Like the private Pearson correlation, one can define an error-based metric that quantifies the distance between the public and private values as mean squared error:

$$MSE_{\text{private}} = \frac{1}{N}\sum_{i=1}^{N}\left(x_{i;\,\text{private}} - x_{i;\,\text{public}}\right)^2$$

Table 5.1: Privacy metrics

Privacy metric	Protection	Approach	Interpretation	Higher Privacy
k-anonymity (Sweeney, 2002b)	Record identification	Blend with other records on identifiable attributes (QID).	Higher values imply higher privacy protection.	High values
t-closeness (Li et al., 2007)	Attribute disclosure	Measures the amount of leaked information (semantic and degree of belief disclosure).	Smaller values imply higher privacy protection.	Low values
Pearson correlation (Kim et al., 2011)	Attribute disclosure	Measures the similarity of original and obfuscated record values.	Higher values imply less privacy protection.	Low Values
Mean Squared Error Asis. (MSE Asis.)	Attribute disclosure	Measures the similarity of original and obfuscated record values.	Higher values imply less privacy protection. Comparative.	Low Values
Mean Squared Error Pred. (MSE Pred.) (Ganti et al., 2008)	Inferential disclosure	Assess an adversary's ability to reconstruct private values from the public dataset.	Low values indicate less privacy protection.	High Values

A further privacy issue in ML is inferential disclosure referring to the indirect inference of sensitive attribute values by reconstructing the original value (Elliot et al., 2005). Inferential disclosure poses a threat to individuals outside the public dataset (Crawford & Schultz, 2014). To quantify the danger of inferential disclosure a ML algorithm is trained on the public dataset to predict the values of the PA attribute and compare this prediction $\hat{x}_{i;\,\text{public}}$ with the private values, $x_{i;\,\text{private}}$ using the mean squared error as a discrepancy measure (Ganti et al., 2008).

$$\text{Pred.MSE}_{\text{private}} = \frac{1}{N}\sum_{i=1}^{N}\left(x_{i;\,\text{private}} - \hat{x}_{i;\,\text{public}}\right)^2$$

A larger mean squared error indicates better privacy protection, as the adversary's ability to infer the original value is limited due to a high level of error (Ganti et al., 2008).

The function $h(x)$ denotes an ML model that makes predictions $\hat{y}_i \in \{0,1\}$ regarding the classification of a data record within the data table $x \in DT$. The evaluation of fairness by an ML model involves the consideration of two, or more, distinct groups, denoted as g_i and g_j, which differ in characteristics such as age (e.g., old, and young) or gender, etc. The label for classification is represented by the symbol $Y \in \{0,1\}$. The terms $TPR(A = g_j)$ and $FPR(A = g_j)$ are defined as the group-specific probability of the classifier $h(x)$ outputting the positive class 1 (e.g., granting a financial credit/loan) given the recipient is eligible (correct classification, true-positive) or not giving the benefit despite eligibility (wrong classification, false-negative):

$$TPR(g_j) = P(h(x) = 1 \mid A = g_j, Y = 1) \text{ or } FPR(g_j) = P(h(x) = 1 \mid A = g_j, Y = 0)$$

A fair ML model allocates benefits equally across demographic groups, requiring equal group-specific performance metrics (Haas, 2019), e.g., false-positive-parity (FPP) and true-positive parity (TPP):

$$TPP := TPR(g_j) = TPR(g_i) \quad \text{or} \quad FPP := FPR(g_j) = FPR(g_i)$$

To measure how fair an ML model is, the deviation from the optimal "fair" situation can be written as:

$$Unfairness(FPR) := 1 - \frac{\min\{FPR(g_i), FPR(g_j)\}}{\max\{FPR(g_i), FPR(g_j)\}} \tag{5.1}$$

The above Equation 5.1 $Unfairness(.)$ measures thus FPP through FPR, but also can be changed to encompass other ML fairness notions, e.g., by replacing FPR with the false-negative rate (FNR) obtaining false-negative-parity (FNP).

5.3 Experiment

We conducted a computational experiment reproducing key steps in a data science pipeline, such as training and evaluation, which are preceded by different PPDTs. PPDTs are benchmarked against predictions from ML models trained on non-anonymized data. Instead of reusing the optimal hyperparameters from the unprotected data on the protected data, models' hyperparameters were independently tuned on both protected and unprotected data, using the F1 score as the optimization criterion. This approach ensures a balanced evaluation and allows for the causal attribution of observed performance losses to the PPDTs.

To enhance the robustness and reliability of the computational experiment, we employed a *resampling strategy* to determine how PPDTs affect fairness averaged across multiple runs. This approach, enables us to identify potential biases and discrimination that may arise from different PPDTs while also controlling for differences between PPDTs that occur simply by chance: due to (1) randomness in the PPDT (e.g., noise addition) or (2) in the ML method being used (e.g., random initialization of NN). We generated $B = 100$ train and test datasets, each with a different distribution of privacy and discrimination-protected attributes (k-anonymity and various levels of disclosure protection). Random sampling with replacement is used to construct resamples for the benchmark, where each sample id denoted by $b = 1, \ldots, B$. Then, 70% of the resamples (bootstraps) are used for training and 30% for testing; doing so always guarantees a constant number of records in the test data to ensure comparability. We evaluated several privacy protection alternatives using four benchmark datasets: German Credit and Adult used in previous literature (Friedler et al., 2019; Haas, 2019) as well as Bank and Diabetes as new propositions. For each dataset, we specified quasi-identifiers, a label, and QID attributes as inputs to the k-anonymity algorithm (Table 5.2). Note, that the attributes in Table 5.2 are written as named in the original dataset i.e. Female is categorical and thus also includes male e.g. as a quasi-identifier. Two binary group attributes ("SensAttr1" and "SensAttr2") are used to evaluate the fairness of each numerical attributes are used two numerical to compute privacy metrics (see Table 5.2). Overall, the evaluation procedure is as follows leading to a total of 52.800 evaluated ML models h :

Listing 5.1: Simulation Study.

1. DATA: Input data
2. DATA: Quasi_identifier QIDs
3. FOR repetition $b \in \{1, \ldots, 100\}$ DO
4. Ids: Sample 1000 rows from data without replacement.
5. train_data DS_{train} : take 700 rows Ids;
6. test data DS_{test}: take the remaining 300 rows from Ids not in train_data
7. FOR PPDT $a \in (onlykano,\ shuffle, noise, microaggregation, patectgan)$ DO
8. IF $a \neq (patectgan)$
9. FOR k $\in (5, 10, 20, 30)$ DO
10. Generalize QIDs such that the size of equivalence classes $\geq$ k and return $DS_{train;public;a,k}$;
11. FOR PPDT parameter $p \in P_a$ DO
12. IF $a = onlykano$ SKIP LOOP
13. Privacy enhance DS_{train} using PPDT method = a and parameter p and save $DS_{train,public;a,k,p}$;
14. END DO
15. END DO
16. ELSE IF $a = patectgan$ THEN synthesize FOR EACH $p \in P_a$ and save result;
17. END DO
18. FOR EACH saved result
19. Load saved result;
20. Compute privacy metrics;
21. FOR $ml\_method \in \{logreg, nnet, rf)$ DO
22. Maximize F1 score on h_{public} using ml_method on $DS_{train,public;a,k,p}$;
23. Compute performance and fairness of h_{public} on $DS_{test,\ private}$;
24. Save results;
25. END DO
26. END EACH
27. END DO
28. FOR $ml\_method \in \{logreg, nnet, rf)$ DO

29. Maximize F1 score on $h_{private}$ using ml_method on $DS_{train,private}$;
30. Compute performance and fairness of $h_{private}$ on $DS_{test,\ private}$;
31. Save results;
32. END DO
33. END DO

Microaggregation and noise addition were performed with the R package sdcMicro. For the synthetic PATE-CTGAN the OpenDP library is used. K-anonymization is based on the authors implementation of datafly (Sweeney, 1998). The caret package in R is used to build a predictive analytics pipeline. The parameterization of PPDT was for noise addition with parameter set for % of added noise $P_{noise} = \{5, 10, 20, 30\}$, Microaggregation with cluster size $P_{mirco} = \{5, 10, 20, 30\}$, and epsilon for differential private SDG in PATE-CTGAN $P_{Patectgan} = \{0.5, 1.5, 3\}$. Shuffling has one dummy parameter $P_{shuffle} = 1$. Original data is also trained for and no privacy-preservation used. The hyperparameters of the ML model are chosen based on 5-fold cross-validation with a grid size of ten evaluation points, according to the defaults of caret. The ML models are logistic regression (logreg), neural network (nnet), and random forest (rf). All simulations are written in R. For the simulation, it is ensured that the same bootstrap samples are always used when comparing PPDTs. The computations were performed on a small cluster of five high-performance workstations as significant computation time was necessary.

Multiple assessments are used to examine the characteristics of the privacy protection methods:

- The *predictive performance* of the ML model, $h_{private}$, is measured on an independent test dataset. This model is trained on the *non-anonymized private* dataset, and its performance is compared to a baseline ML model, h_{public}, which is trained on the *anonymized public* dataset before dissemination. The performance differences between the two models estimate the utility losses when using a privacy-protected dataset. We assess predictive performance using well-accepted criteria such as accuracy, recall, precision, and the F1 score.
- The outcome of *fairness* is evaluated by calculating the relative difference between the fairness of h_{public} and $h_{private}$. To assess the impact of privacy protection on different demographic groups, we use Equation 5.1.
- t-closeness, Pearson correlation, and mean error quantify *privacy protection*. The latter metric uses a one-layer neural network of size num_vars and weight decay to assess the prospect of reconstructing sensitive numerical attributes. If this quantity is measured

using unseen test data we refer to it as pred_mse. Asis_cor and asis_mse quantifies the similarity between the private and public dataset after applying PPDT. To calculate t-closeness, we adopt the suggestion of Li et al. (2007) using ordered distance for numeric attributes.

Using the above setup, we conducted the computer experiment. For the results, we report first the effect on overall performance, then the results on fairness and privacy.

Table 5.2: Attributes used for evaluation.

Dataset	**QID**	**Sensitive Attribute**				**Label**
		Fairness		**Privacy**		
		1	2	1	2	
Adult (a)	Age, Female, race	Age < 30	Race = 'Caucasian'	capital_gain	capital_loss	Salary
Bank (b)	age, marital	Age < 30	marital	balance	Duration	Y
Credit (c)	Age, Female, Foreign	Age < 25	Female	Amount	Duration	GOOD
Diabetes (d)	age, gender, race	age	gender	time_in_hospital	num_medications	readmitted

5.4 Results

5.4.1 Analysis.

We analyze the effects of PPDT on performance using a *hierarchical linear model* (HLM) with PPDT, ML method, and dataset as fixed effects to obtain estimates. Random effects were modelled as $\delta_{d,b} \sim \mathcal{N}(0, \sigma_d)$ which is a random effect for each dataset, $d \in 1,\ldots,4$, and bootstrap sample, $b \in 1,\ldots,100$, accounting for the nested structure of bootstrap resample within datasets. Statistical significance was determined using Satterthwaite's approximation method for the HLM.

5.4.2 Performance

First benchmark results for performance as a dependent variable are analyzed. The HLM in Table 5.3 shows that PPDTs consistently reduce performance compared to ML models trained on protected data across all performance measures. These reductions are negative and statistically significant for most combinations of PPDT and performance metrics. In addition, the performance losses are, in most cases, practically significant, amounting to at least three to four percentage points. These are substantial trade-offs for ML models deployed in high-volume predictions or safety-critical domains. To visualize the performance loss through PPDT violin plots were used (Figure 5.1).

When examining specific PPDTs in Table 5.3 in terms of performance losses several patterns emerge. Firstly, microaggregation, noise addition, and only-k-anonymity – all belonging to the category of perturbative methods – show the smallest performance loss in accuracy. Those PPDTS reducing performance by approximately four percentage points on average. In contrast, PATE-CTGAN, a synthetic one, and random shuffling introduce larger losses in accuracy. The pattern remains consistent across F1 and Recall metrics. Overall, the performance losses vary less within the first three PPDTs and the last two PPDTs. Those PPDTs are *clustering* as the performance losses of k-anonymity alone is close to those for microaggregation/noise addition (which combines k-anonymity with attribute protection). This observation suggests that once datasets are protected against identity disclosure, adding attribute disclosure protection comes with minimal further performance cost on average. One interpretation is that the feature importance (FI) of these features is low. FI measures the predictive contribution of features to overall performance by shuffling (i.e. randomizing and permuting) the feature's records while keeping other features constant. The difference in performance between original/shuffled then quantifies the importance of that feature. PPDTs are designed to retain information, whereas shuffling nullifies relationship between a target and feature. PPDTs are also applied to multiple features simultaneously, while shuffling in FI to a single feature. The only-k-anonymity method is applied solely to quasi-identifiers. The small performance loss difference between noise addition and microaggregation may be explained by two interpretations based on FI. First, the features protected by these PPDTs may have minimal predictive contribution to the target variable. Alternatively, noise addition and microaggregation may only marginally affect the original relationships in the dataset – which is a second interpretation of the results. Shuffling, which completely nullifies the original relationships and represents the worst-case scenario, shows a two to three times larger performance loss compared to microaggregation and noise addition. This suggests that microaggregation and noise addition indeed introduce only minor

Table 5.3: Performance (HLM regression)

	Accuracy			F1			Recall			Precision		
	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p
1	-4	39	***	-4	20	***	-6	23	***	-1	5	***
2	-4	34	***	-4	18	***	-6	21	***	-0.3	3	**
3	-4	39	***	-4	20	***	-6	23	***	-1	5	***
4	-15	82	***	-15	52	***	-19	47	***	-7	42	***
5	-10	85	***	-9	45	***	-10	35	***	-7	54	***
ML												
	Accuracy			*F1*			*Recall*			*Precision*		
	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p
nn	2	34	***	2	27	***	4	38	***	-0.2	-6	***
rf	2	37	***	3	30	***	5	42	***	-0.1	-3	**
Dataset												
	Accuracy			*F1*			*Recall*			*Precision*		
	Est.	|t|	P	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p
b	2	12	***	1	4	***	1	4	***	1	5	***
c	.	3	**	-5	20	***	-4	11	***	-1	5	***
d	1	7	***	-1	2	*	-1	3	**	0	1	ns

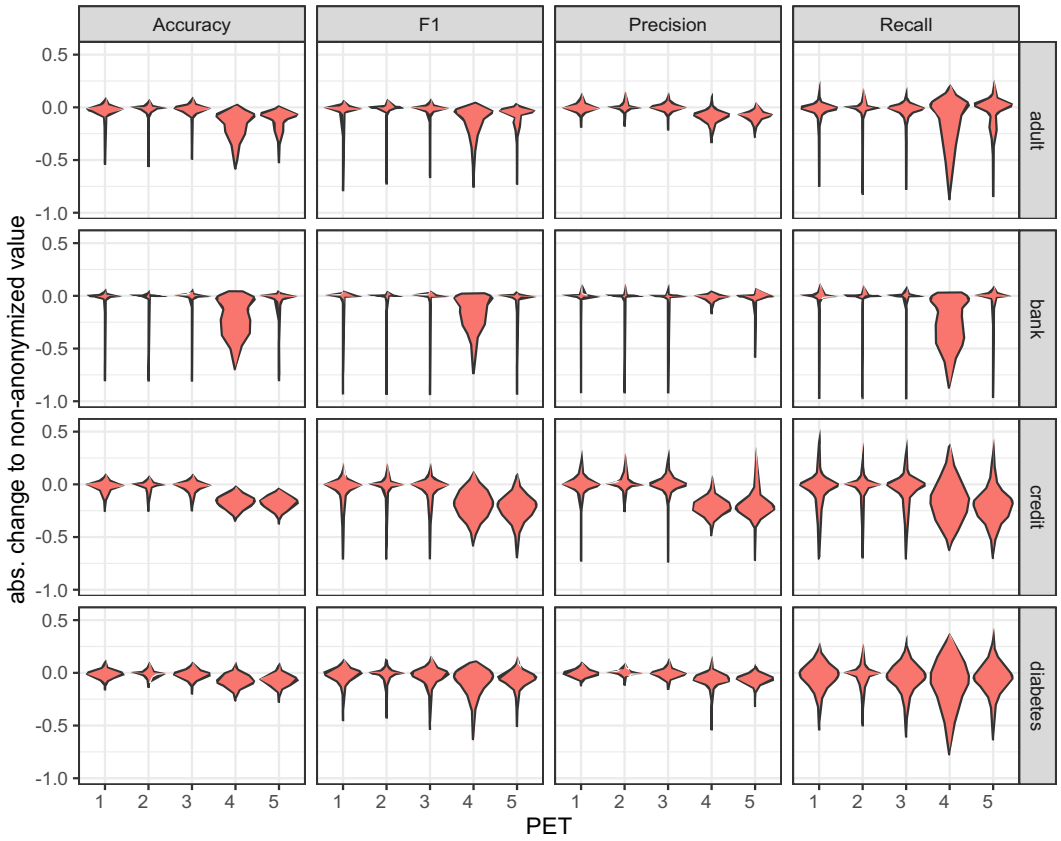
distortions to the original dataset. Therefore, building on the second interpretation of the results, the conclusion that microaggregation and noise addition incur minimal costs to predictive performance once k-anonymity is established remains.

To complement the analysis of k-anonymity versus microaggregation/noise addition we conducted a statistical test on the distribution of performance losses. For this test, we performed post-hoc 32 pairwise Kolmogorov-Smirnov tests to compare only-k-anonymity with the PPDTS methods (micraggregation and noise addition) across all possible combinations of four datasets and four performance metrics ($2 \times 4 \times 4$). To address the issue of multiple comparisons ($m = 32$), we applied an adjusted significance level of a/m, where a is set at the 10% significance level. Out of these 32 tests, 29 yielded non-significant results, suggesting that the distribution of only-k-anonymity does not significantly differ from the tested PPDTS methods. In three cases ‘Acc, noise, diabetes’, ‘Prec, microMean, diabetes’, and ‘Prec, noise, diabetes’ the result

was significant. For PATE-CTGAN and shuffling, we did a separate KS-test which also statistically significant and which we interpret as being consistent with the numerical size of the average scores form the HLM. Reiterating on the three cases of significant results comprising the perturbative methods we conduced additional tests. We conducted a post-hoc equivalence test using Two One-Sided Tests (TOST) with a Welch Two Sample unpaired t-test (Lakens et al., 2018). The equivalence bound was set symmetrically $\Delta_U = |\Delta_L| = 0.015$ (one and half percentage point). These bounds were determined after analyzing typical variation ranges in ML benchmarks on Kaggle and our own simulation analysis, where the natural variation in the subset of the original unaltered data standard deviations within the Perf. Metric, dataset and ML method ranged around 1.5 to 3 percentage points. We selected symmetric bounds as we considered deviations in either direction equally meaningful. This bound tests if the difference between these noise addition/microaggregation and k-anonymity is larger than one percentage points, which forms the H_0. In those 32 PPDTs the combination 'Rec, mircoMean, bank' was the equivalence test non-significant, (t = -1.5, df =8251.83, p = 0.06). These results suggest that there is no remarkable loss of model performance in favor of providing added privacy protection using perturbative methods.

Next, we compare the *performance metrics* in Table 5.3. Between the different metrics the level of performance loss for these metrics differs considerably at least by some percentage points. Among all performance metrics Recall is the most affected by PPDTs, with largest loss of up to 24 percentage points and the smallest close to six percentage points. The large effect on recall suggests that PPDTs can substantially affect the model's ability to identify true-positives, in particular for PATE-CTGAN and shuffling. At the same time, it is noteworthy that PPDTs have the smallest impact on precision, particularly for microaggregation, k-anonymity, and noise addition, where the loss is nearly zero and for the remaining PPDTs not statistically different from zero. In general, precision and recall in ML form a trade-off such that those objectives cannot be optimized simultaneously. Consequently, the notable phenomenon is not the inverse relationship between recall and precision, but rather the consistent decrease of recall across all PPDTs. All ML models were optimized for the F1 metric and the same metric is the less affected performance metric after Precision in the evaluated models and datasets but Precision still has smaller performance losses than the F1 score. This is particularly remarkable as all ML models were tuned using grid search specifically for the optimization of the F1 score. Since the F1 score represent the primal optimization criterion it is notable that precision but not F1 is less affected by PPDT. In our analysis, we observed that the F1 score was optimized by increasing precision in anonymized datasets, as the differences to the private datasets are small or close to

Figure 5.1: Performance (violin plot).



PPDTs. 1: microaggregation. 2: only k-anonymization. 3: noise. 4 PATE-CTGAN . 5 shuffling

zero, but at the expense of detecting positive class labels leading to a significant drop in recall. Also, PPDTs impact performance metrics unevenly, particularly the components of the F1 score. Beyond reducing overall performance, PPDTs alter the balance between precision and recall (as the performance loss for recall is larger than for precisions). The impact is not proportional across all metrics, making comprehensive evaluation essential for understanding the performance trade-offs involved with privacy. Using PPDT therefore implies altering economic incentives when considering the costs of false-positives, which precision mainly optimizes for, and false-negatives, which recall mainly optimizes for. PPDTs can change economic trade-offs and lead to new optimality criteria for finding best actions in the decision space when grounding decision-making on the privacy-preserved data instead the original data.

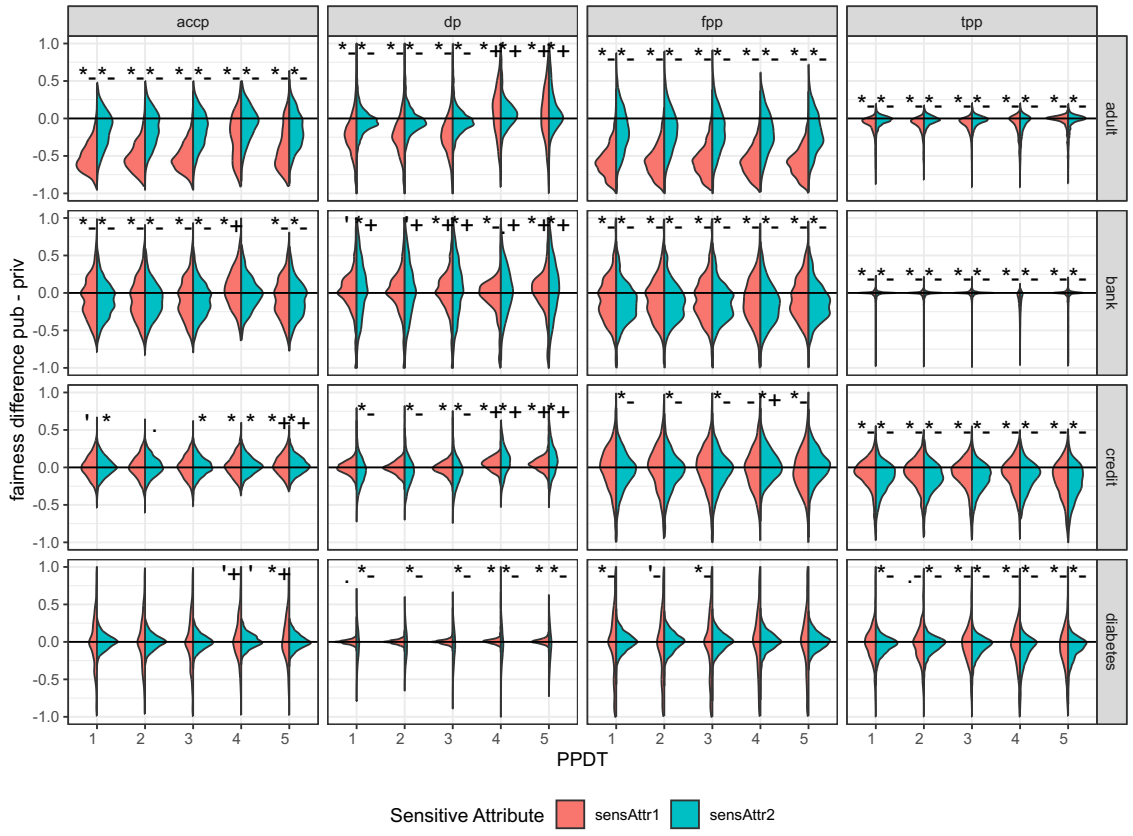
Comparing the ML models in Table 5.3, both the neural network and random forest showed improved accuracy over the logistic regression (which is the omitted category in the HLM). The random forest demonstrated the largest improvement, while the neural network performed similar well. This indicates that complex ML techniques performed on average better using the privacy-preserved data compared to logistic regression. As the overall effect size of the two

ML models has smaller magnitude than the one of the PPDTs the total effect of applying a complex ML model on PPDT data is negative. Hence, complex ML models fit better to privacy-preserved data than less complex ML simpler ML only can partially outweighed the loss of accuracy introduced through the PPDTs. As less complex ML models are more interpretable, PPDT increase the trade-off between performance and interpretability. Concerning the heterogeneity of the four datasets, only the diabetes dataset (where the adult dataset is the reference level) demonstrated a significant increase in performance, both statistically and in magnitude. The other datasets exhibited a small but statistically insignificant decreases.

5.4.3 Fairness

To assess the impact of PPDTs on the fairness of the algorithmic outcomes, we computed fairness metrics for two groups on each dataset, following the setup of Table 5.2. Given the multiple outcome measures per dataset, including two group fairness measures (red sensAttr1, blue sensAttr2), we employed two-sided t-tests. To account for multiple comparisons, we adjusted the p-values using the Bonferroni correction. We report p-values as significant if they are less than the conventional significance level divided by the number of comparisons (160 in total). The results are depicted in Figure 5.2. Significance levels indicated by * $p < 0.001$, . $p<0.01$, ‘ < 0.05. Figure 5.2 displays that PPDTs exert heterogeneous effects on fairness across different metrics and datasets. The directional impact varies significantly — when PPDTs meaningfully affect fairness, these effects can be either positive or negative. Notably, PPDTs 1, 2, and 3 consistently diminish fairness across all examined datasets and metrics. Conversely, PPDTs 4 and 5 improve fairness in some contexts but not universally. The visual impression is also supported by numerical quantities. For the FPP fairness metric, our findings at the 5% significance level indicate that PPDTs decrease fairness by at least 2.5 and 5 percentage points in 67.5% and 57.5% of the runs in the benchmark study, respectively, compared to using no PPDT. For ACCP, there is a significant fairness reduction in 45% of cases for both thresholds and for TPP, it is 55% and 92%. This suggests that the effect of PPDTs on fairness is 1) not unidirectional and 2) often considerably negative.

Figure 5.2. Fairness Metrics (violinplots)



PPDTs. 1: microaggregation. 2: only k-anonymization. 3: noise. 4 PATE-CTGAN. 5 shuffling

Since fairness demographics often overlap with quasi-identifiers (e.g., gender, ethnicity, age), the relative fairness changes of k-anonymity versus perturbative/synthetic PPDTs were analyzed. Investigating the implications of the overlap is important as k-anonymity model could disturb fairness-related attributes, as these are also quasi-identifiers (QID) (see Table 5.2). We compared therefore the k-anonymity with PPDT focusing on attribute disclosure and synthetic data using HLM regression, including sensAttrs as fixed effects (see Table 5.4). We found that microaggregation (referred to as 1 in Figure 5.2) and noise addition (referred to as 3 in Figure 5.2) had no statistical significant different effect than those introduced by only k-anonymity. Noise addition had no statistical significant effect on all fairness metrics but the TPP metric, with the effect being relatively small in size. Microaggregation showed no significant effect on all fairness metrics, potentially suggesting it introduces no additional unfairness in the dataset beyond that introduced by k-anonymity. To further elaborate on this result, we conducted the post-hoc equivalence test TOST as for the analysis of performance. In the fairness analysis we set to $\Delta_U = |\Delta_L| = 0.01$ initally. Performing this TOST test for the TPP metric, with null hypothesis of extremer effect, rejects this null hypothesis (t= -4.7, df = 15907.19, $p < 0.01$) and

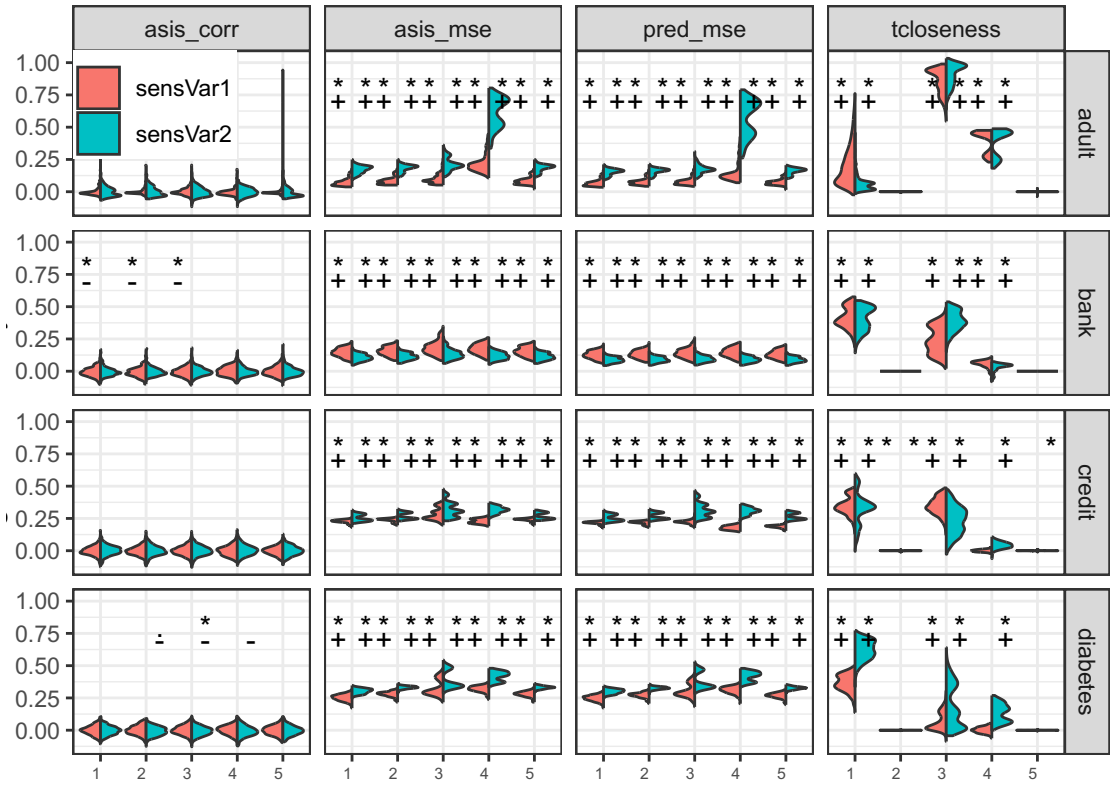
similar equivalent result for DP (t= 4.85, df = 15196.9, p < 0.01). In contrast, performing the same test failed to reject the null hypothesis for FPP (t= 1.58, df = 15373.63, p = 0.06) and ACCP (t=0.54, df = 15332.62, p =0.3). Therefore, statistically microaggregation has no practical difference from k-anonymity in the range of +/- 1-percentage point in the simulation for the DP and TPP metric, whereas for FPP and ACCP the equivalence hypothesis did not hold. Since the ACCP and FPP metrics share the number of false-positives in their calculation, the results may be dependent. Overall, the difference is small on average. Increasing the equivalence bound to 2 and 1.5 percentage points, respectively, suggests practical significance between the results for the FPP and ACCP metrics.

The analysis of fairness metrics in Figure 5.2 across multiple datasets reveals three distinct patterns in how PPDT creates algorithmic bias. This first pattern exhibits a zero-sum fairness distribution, wherein the frequency of fairness improvements for the affected demographic group is equivalent to the frequency of fairness decreases, resulting in balanced probabilistic outcome across the benchmark experiment where the averaged effect is close to zero. This pattern is for instance visible in the diabetes dataset for group 1 and group 2. This pattern occurs in two variations. The fist variation has small risk meaning that the spread of a gain/loss in fairness is relatively modest. We refer to a modest variation as it appears in Figure 5.2 when violin plot is approximately symmetrically shaped around zero and its edges do stretch not too much to the edges of the Figure 5.2 as do other violinplots. For example, Group 1 for the DP metric or group 2 for the ACCP metric in the diabetes dataset or group 1 and group 2 in the bank dataset may be best described by this low-risk fair gambling pattern. The second variation is the high-risk one. This variation is as well symmetrically shaped around zero but the plot stretches more to the edges of the Figure 5.2 and to the theoretical upper levels of the metric that are computed. Group 1 in diabetes dataset for ACCP and FPP metric for instance may be characterized by following this pattern. Group 2 in the bank dataset for the DP metric has also a relatively high-risk, meaning high variations in this fairness metric, whereas for group 1 this variation is much smaller. On average this groups are treated equally in terms of fairness losses as their expected value in the simulation are numerically close to each other.

Table 5.4. Fairness (HLM regression)

PPDT Fairness												
	ACCP			DPP			FPP			TPP		
	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p
0	-32	-32	***	-11	-14	***	-40	-28	***	-8	-18	***
1	0	0	ns	0.1	1	ns	0.1	0.4	ns	0.2	1	ns
3	0	0	ns	0.2	-1	ns	0.2	1	ns	-0.4	-3	**
4	7	2	ns	6	2	ns	0.1	0	ns	-4	-2	*
5	3	10	***	8	26	***	1	3	**	-2	-9	***
ML												
	ACCP			DPP			FPP			TPP		
	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p	Est.	|t|	p
nn	-4	-24	***	2	10	***	-1	-4	***	4	33	***
rf	-4	-24	***	0.3	2	ns	-1	-4	***	4	29	***
Dataset												
	ACCP			DPP			FPP			TPP		
	Est.	|t|	p	Est.	|t|	p	Est.	|t|	P	Est.	|t|	p
b	21	16	***	13	12	***	25	13	***	4	7	***
c	31	23	***	7	7	***	33	17	***	-6	-10	***
d	29	22	***	7	6	***	33	17	***	0	0	ns

Figure 5.3. Privacy metrics (violinplots)



PPDT. 1: microaggregation. 2: only k-anonymization. 3: noise. 4 PATE-CTGAN. 5 shuffling

The last pattern is the pattern of structurally disadvantaged groups through PPDT. Group 1 for accp and fpp in the adult dataset forms one such example. This group is always treated worser after PPDT. Interestingly, for the same group and dataset there are also occurring combinations of fairness metric and PPDT in the DP metric and PPDT 4 and 5 which increase fairness for the group as measured by the metric. This observation can be denominated as a strong *metric divergence* which is observed, as PPDTs improve some fairness metrics while simultaneously degrading others. These findings show that fairness interventions can have complex effects beyond average metrics, with some groups experiencing higher variability or systematic disadvantages that must be considered in algorithmic fairness evaluations.

5.4.4 Privacy

Finally, we also compare PPDTs based on how much privacy they provide (Table 5.3). For assessing privacy, we use again a similar setup as for the assessment of fairness evaluating two sensitive attributes for which attribute disclosure might be a concern (see Table 5.3). For all metrics except, correlation, an increase means better privacy protection, whereas for correlation, smaller values mean more privacy protection. We find that overall, PPDT mostly provides privacy protection according to the latter criteria. Notably, observed and inferred MSE

are overall similar in magnitude, implying that the sensitive attributes are not easily reconstructed using ML as indicated by keeping a high error on inferred MSE. In a separate analysis (details omitted), we examined the correlation between privacy metrics. We found that asis_corr and asis_mse are practically equivalent. Optimizing one of these metrics also results in a high pred_mse, indicating protection against attribute disclosure by reconstruction attacks, even for individuals outside the public dataset. T-closeness has low correlation and therefore captures distinct privacy aspects.

5.5 Discussion, Implications, and Limitations

5.5.1 Discussion

Previous research on privacy, fairness, and performance either studied them separately or focused on specific algorithms and privacy concepts. Notably, studies on differential private ML methods (Cummings et al., 2019; Bagdasaryan et al., 2019b) and federated learning (Chen et al., 2023) address specific scenarios of privacy protection when data sharing is not for analysis purposes. Other research studied the effect of enforcing fairness first on privacy, conceptualized through the risk of identity disclosure (Chang & Shokri, 2021). Research on privacy-preserving data sharing methods in a sociotechnical framework that considers utility-fairness-privacy holistically specifically for tabular business data is underexplored (Conde et al., 2024). Closest to our research is Carvalho et al. (2023a), who found a trade-off between privacy protection against identity disclosure and performance. Our work differs from this earlier research by focusing on privacy conceptualized by attribute disclosure and its effect on several measures of algorithmic fairness. For certain PPDTs like microaggregation and noise addition, we found no such trade-off on average, which were not included in Carvalho et al. (2023a). For other types like CTGAN, we reproduced Carvalho's finding of a performance trade-off using a different methodological setup and PATE-CTGAN, which used PATE for additional privacy. A very recent work by Liu et al. (2025) compared synthetic data methods for fairness-privacy-utility. They found PateGAN to prioritize privacy over other metrics. Therefore, PateGAN, despite having high privacy guarantees may be a one-sided candidate for synthetic PPDT due to its inferior utility-fairness (Liu et al. 2025). The study of Liu et al. (2025) was published after our benchmark study was completed and we could therefore not consider its findings. However, this study differs in some important regards. Firstly, we report for PATE-CTGAN, an extension of the PateGAN – the latter was used in Liu et al. (2025) -, that is better suited for handling categorical data. Categorical/one-hot-encoded data is a challenge for PateGAN when generating artificial data (Xu et al., 2019). As categorical data is often found

in business analytics for customer data (gender, marital status, sales channel, product categories et.c.) we included Pate-CTGAN in this study. It may be that this algorithm performs relatively better than PateGAN baselines. Furthermore, Liu et al. (2025) did not use resampling, to account for randomness in the procedures, which they also note as a limitation. In this study, we accounted for randomness by taking many resamples thereby complementing earlier research by a different methodological approach. While CTGAN was effective at generating privacy-protected data, its fairness performance varied

Our application of PPDT before studying fairness impacts reflects the perspective of data management departments that are responsible for securely disseminating data. This approach specifically addresses organizational scenarios where data preparation and analysis are separated responsibilities. Overall, we found PPDT result in performance loss. Regarding how this loss distributes over demographic groups, PPDT can affect ML fairness in both directions. Frequently, unfairness worsens, although certain combinations of dataset and sensAttr exhibited improved fairness after privacy enhancement. This improvement may be attributed to the fact that privacy enhancement made individuals more similar, resulting in similar outcomes. Compared to existing literature where higher performance often correlates with increased fairness and vice versa (Friedler et al., 2019; F. Kamiran & Calders, 2012; Haas, 2019), our results suggest this correlation does not necessarily apply to PPDTs: while reducing overall performance, fairness loss appears evenly distributed on average. Furthermore, PPDTs can not only protect sensitive information but also promote fairness in certain resampled or trained ML models, though these effects vary inconsistently across datasets and groups. This implies that in a data science pipeline, preparing microdata for sharing might require multiple iterations of PPDT and subsequent analysis to optimize PPDT not just for privacy but fairness as well, considering its propensity to have *bi-directional* effects. We extended earlier research on PPDT (Carvalho et al. 2023a; Liu et al. 2025) by accounting for randomness in the process and find that trade-offs are not necessarily a stable property. Furthermore, this study compares GAN-based synthetic data with perturbative methods such as noise addition and microaggregation. Results indicate that synthetic data PPDT underperform perturbative methods across multiple predictive performance metrics. While synthetic data has several attributes that make it appealing from a privacy standpoint, this research investigates whether these properties remain when considering fairness and performance in comparison to perturbative methods. To the best of our knowledge and reviewing sufficiently extensive recent literature review the comparison of perturbative methods and synthetic data is novel in the research stream (Conde et al., 2024; Carvalho et al. 2023a). The most effective PPDTs for

supporting model performance are microaggregation, noise addition, and k-anonymity, showing small losses in accuracy and other performance metrics. PATE-CTGAN and random shuffling lead to more substantial performance declines. While PATE-CTGAN and shuffling perform significantly lower than other PPDTs, they are not directly comparable as PATE-CTGAN generates synthetic data while shuffling changes a subset of attributes in the original data. Remarkably, k-anonymity alone achieves performance close to microaggregation/noise addition (which combine k-anonymity with attribute protection). Microaggregation is more effective than noise addition in maintaining overall fairness. This indicates that after datasets are safeguarded against identity disclosure, implementing attribute disclosure protection via microaggregation incurs only a slight increase in performance and fairness costs.

In the results, we identified two patterns of the effect of PPDT on fairness. These patterns can be distinguished by the expected value and variance of the fairness difference between private and public data (see also Table 5.5). When interpreting these patterns, one must critically examine whether the „fair gambling“ pattern truly represents fairness. Notably, the high-risk variation exhibits substantial outcome variability despite statistical balance, raising questions about whether it reflects genuine algorithmic fairness or merely mirrors the underlying data distribution. To clarify this methodological consideration, each fairness metric calculation uses resamples from the original unprocessed data. While both the privacy-enhanced and original datasets experience compositional changes during resampling, these changes occur consistently across both sets (i.e. the same records). The results indicate for some combinations of PPDT, fairness metric and dataset there is variability in fairness metric with this metric taking fair/unfair realizations depending on changes of the composition in the test dataset. As a consequence, those members of the group sometimes receiving an unfair treatment. Stochastically these unfair treatments average out taking many of those test datasets but nevertheless appear to exist temporarily. This means that deployed ML systems can exhibit unfair treatment temporarily as well for certain subgroups. While this variability may be a natural phenomenon, it is still noticeable that this variability is higher for certain groups and lower for others, as indicated by our benchmark results (e.g. for accp, dp, fpp in the diabetes dataset). Therefore, using PPDT can increase the temporarily unfairness of a deployed ML system for certain groups while other groups have less extreme fair/unfair results. However, the results show that simply changing the PPDT does not change this behavior as the large variability is consistently seen across all PPDT.

Table 5.5 Identified fairness harms.

Name	**Characteristic**	**Implication**
Systematic harm	PPDTs consistently disadvantage specific demographic groups	Alternative PPDTs should be evaluated to avoid systematic bias before implementation.
Stochastic harm (i.e., "fair" gambling)	PPDTs impact groups randomly with neutral average effect but variable outcomes	Extensive testing of PPDT and ML hyperparameters across metrics and demographics required. Continuous post-deployment monitoring essential to prevent fairness violations.

5.5.2 Managerial Implications

Data within organizations involves multiple stakeholders each with distinct roles and responsibilities. Primarily, data management serves as the provider of data, with less frequent involvement in data analysis. Due to legal restrictions and privacy concerns, data managers are obliged to protect the privacy of the data they keep, even when sharing it within the organization. This raises the question: what are the *downstream implications* of applying PPDTs before distributing it to other stakeholders, such as data scientists and managers? The objective of this research was to examine how PPDT affects the outcomes privacy, fairness and predictive performance each operationalized through several metrics. The research objective bears significant practical relevance. Legally, organizations are mandated to implement privacy protection strategies when utilizing customer data. At the same time, organizations using ML and AI systems are required through legislation such as the EU AI Act[5] to conduct impact assessments of these systems they use. Therefore, pressure on organizations is increasing to conduct ML and AI responsibly (van der Aalst et al., 2017).

To ensure this, managers should establish protocols that allow for regular cycles of PPDT review and fairness evaluation, potentially by having closer ties between data management and data science/analysis. First, *practitioners can utilize the evaluation procedure* in Algorithm 1 to inspect their systems' behavior. Not considering the stochasticity of PPDT and ML

[5] p. 13-24, https://ec.europa.eu/newsroom/dae/redirection/document/75792

algorithms concerning their outcomes may lead to flawed conclusions about the system characteristics, which do not stand up to analysis, as showed in our study. Our findings are strictly applicable to technical privacy and ML fairness. Behaviorally, perceived fairness and privacy could also depend on PPDT transparency and trustworthiness. Though some PPDT score well on privacy metrics, they do not necessarily achieve perceived privacy. For example, the simplicity and transparency of methods like noise addition may engender more trust than the complex, black-box PATE − even if the latter promises high privacy protection by being synthetic data. Since no PPDT outperformed all objectives (ML utility, fairness, and privacy), in-practice data dissemination for specific purposes may be necessary. For instance, enterprises can share a dataset with strong privacy protection and high descriptive attribute preservation using PATE-CTGAN for prototyping and idea testing. Then, they can fine-tune an ML model on a dataset protected by a PPDT, such as microaggregation, with smaller utility loss in predictive performance. This restricted setting (e.g., under supervision) requires less privacy protection.

5.5.3 Limitations

Some limitations should be noted. First, the PPDT relative rankings may differ for other datasets. Consequently, practitioners should consider our research as a comprehensive guidebook for the preparation of their data distribution. Second, our research focuses solely on numerical tabular data, needing the adaptation of many evaluation procedures for other data types. However, in business, most data is stored in tabular form. Third, the effect on fairness for sensAttr1 could be an artifact of the age attribute's categorization. We have set this attribute to 25 years in the credit data set and to 30 years in the remaining datasets to form young/old age groups. Fourthly, results are limited by the fact that the studied datasets, ML methods, and privacy protection methods are only a small subset of all combinations of these methods. Hence, findings are limited with respect to generalizability. This is a limitation inherent to this type of studies due to computational restrictions. It is still important to conduct this type of studies because otherwise, important phenomena relevant to the sociotechnical discipline of IS remains underexplored. We use recognized benchmark datasets, and such, our results contribute to the cumulative tradition on research on these datasets. Finally, while research has explored the perceived fairness of technical fairness metrics, the relationship between technical privacy metrics and perceived privacy appears underexplored. Because the ethical ML framework is behaviorally motivated, considering how the ethical ML pipeline and perceived privacy of PPDT affect customer behavior is important.

5.5.4 Future Research

Some interesting avenues for further research would be the perceived privacy of PPDT and privacy metrics, as well as empirical experiments examining behavioral outcomes such as voice for trade-offs between privacy and technical fairness. On a technical side, it would also be interesting to see if there is a trade-off concerning the decreased transparency of ML models after using PPDT

III. Machine Learning Systems for Optimization and Management

ML is frequently utilized in business for tasks such as classification and regression, addressing prediction challenges, and automating tasks previously handled by humans. However, ML can also be leveraged to facilitate decision-making and provide actionable policies. Prescriptive Analytics Systems (PAS) are sophisticated tools that enhance organizational decision-making by offering actionable insights derived from data analysis (Wissuchek & Zschech, 2024). These systems are typically employed by businesses and goal-oriented organizations to optimize processes, improve operational efficiency, and drive strategic decisions. This chapter includes four articles on innovative approaches to using ML for decision support and prescriptive analytics in specific organizational contexts.

The first article, "Prescriptive Analytics in Procurement: Reducing Process Costs" describes a new approach to the procurement literature (Bienhaus & Haddud, 2018; Spreitzenbarth et al., 2024) addressing process costs in supplier search. The second article, "A multi-objective Particle Swarm Optimization Framework for Operations Management" describes how ML can be used for tasks that are "wicked", in a sense, that they do not only lack of label and have several conflicting objectives to qualify the solution space. The proposed approach is new to the operations management literature (Wissuchek & Zschech, 2024). The second article uses the application of the first as a case study but the described method can be applied to a wide range of problem, not only procurement.

The third article "Robo-Advisory and Algorithmic Trading via Evolutionary Discretization and Rule-Mining" and fourth article "Bike-Sharing Station Placement: Spatial Analysis and Data Mining of Network Design Characteristics" describe further applications of ML to specific organizational optimization and management problems.

6. Prescriptive Analytics in Procurement Reducing Process Costs

Julian Sengewald (TU Dortmund)

Richard Lackes (TU Dortmund)

Abstract. In obtaining low-cost goods, the indirect expenses associated with sourcing suppliers can be substantial compared to the potential advantages of lower direct purchase costs. We addressed this problem as an 'exploration' vs. 'exploitation' trade-off. The proposed methodology uses a Bayesian technique to learn a stochastically optimal sourcing strategy from quotation data directly. We illustrate our approach using real quotation data for the procurement of electronic resistors (n=201,187). Rather than making optimal predictions, we concentrate on making optimal decisions. In doing so, we offered a significant improvement in purchase and procurement process costs. Our model is also more robust to prediction errors.

6.1 Introduction

Employees in organizations often spend a considerable amount of time on tasks with uncertain outcomes. A particular context where such a problem exists is supplier search in procurement. In procurement, a purchasing agent must search for the best supplier source for the company. To find the best supplier, the purchasing agent must first survey the supplier market and obtain a price quotation from each supplier for the specific purchase. However, the problem for the purchasing manager is that procurement prices are unknown before identifying, approaching, and negotiating with a supplier. In addition, the cost of acquiring a price quotation is spent ex-ante. On the other hand, the potential cost reductions associated with receiving a lower-priced quotation are contingent on the unknown price and are only discovered ex-post. To summarize, finding a better supplier quotation is often not guaranteed.

Another significant aspect of supplier search is that identifying a supplier source takes hours of investigation, supplier verification, and evaluation. Hence, procurement done exclusively and extensively by humans makes supplier search time-consuming. While the primary aim of every purchasing manager is to minimize direct purchase costs, any savings from acquiring goods at a lower price therefore must be balanced against increased procurement process costs Boer et al. (Boer et al., 2002). Traditionally, purchasing managers utilize a curated list of a few vendors to acquire quotations or limit the number of obtained quotations, especially for low-cost items. However, a fixed limit may not be optimal.

The trade-off between learning new information and using the learned information is often called the "exploration" vs. "exploitation" dilemma (Sutton & Barto, 2018). This trade-off is a

main question of research on optimal stopping, reinforcement learning, and bandit algorithms Sutton & Barto (Sutton & Barto, 2018). Ideas from this type of research have been successfully adapted to business problems such as optimal pricing experiments Ferreira et al. (Ferreira et al., 2018), order release decisions (Schneckenreither & Haeussler, 2019), production scheduling Gabel & Riedmiller (Gabel & Riedmiller, 2011), or inventory management (De Moor et al., 2022) – each area developing unique solutions for the specific settings in these applications.

The "exploration vs. exploitation dilemma" is also present in procurement. In addition, there is the problem of the relatively high exploration cost of obtaining price quotations from the supplier. Supplier search in procurement can therefore be reframed as a problem of optimal stopping. An analytics solution that solves this problem can help purchasing managers decide how many resources (e.g., person-hours) should be allocated to a specific procurement task. By doing so, we recognize that much of procurement involves certain work steps that cannot be further automated and that targeted resource allocation is required. Such analytics problems can be seen as prescriptive analytics problems Bertsimas & Kallus (Bertsimas & Kallus 2020). To the best of our knowledge, no previous study has considered procurement automation a problem of optimal stopping. The purpose of this study is to address this problem. Therefore, we ask the following research question:

***RQ**: How can we help procurement managers to balance direct purchase and overall procurement process costs?*

To answer this research question, we investigated a practice-motivated problem in procurement. More specifically, we examined the problem of obtaining low-cost goods electronic resistors, where the indirect costs related to selecting suppliers (procurement process costs) are often substantial in proportion to the benefits of lower direct purchasing costs. Electronic resistors can be found in every electronic device (e.g., washing machines, lighting systems). With prices typically ranging between a few cents and up to a few euros, resistors are relatively cheap compared to the devices they are used in. Resistors come in various materials (e.g., carbon, ceramic), types (e.g., axial, surface mounted), and sizes. Purchase departments must therefore manage a sizeable quantity of different items, often from separate suppliers. The study grew out of a continuing collaboration with a German SME (small and medium-sized businesses) whose management identified the need to improve management and control of sourcing and procurement processes.

We investigated a significant issue within supply chain automation, a classic research problem Toorajipour et al. (2021). We were particularly interested in algorithmic characteristics that

balance decreasing direct purchase costs with increasing process expenses. For this, we calculated the expected discount of searching for a lower-cost supplier offer based on the current best available offer. We also studied a Bayesian strategy for improving machine learning estimates based on actual supplier price quotations. Our proposed technique considers model uncertainty and its impact on decision-making to generate sound prescriptive predictions. Our study contributes to the information systems literature by proposing a novel prescriptive machine learning method with impactful implications for supply chain practitioners.

6.2 Related Literature

To date, several studies have investigated procurement automation (see Table 6.1). The first step that can be automated is supplier discovery (e.g., by mining online news documents) and the collection of price offers Cui et al. (2022). After suppliers have been identified, the best supplier has to be selected among a pool of candidates, for which different multi-criteria decision-making techniques exist, when selection criteria can be explicitly stated (Scott et al. 2015). Alternatively, historical data could be used to infer those selection criteria automatically Wu (2009). Another body of research helped purchasing managers determine the optimal ordering frequency/quantity (Scott et al. 2015). Automation in supplier negotiation is also a topic Cui et al., 022).

Table 6.1: Literature Review

	Step being automated				
Reference	Identification	Selection	Negotiation	Spent optimization	Finding
(Wei et al., 2013)	✓	✗	✗	✗	• Text and link mining techniques can be effectively used for discovering suppliers from online news documents.
(Cui et al., 2021)	✓	✗	✗	✗	• If chatbots collect supplier offers, they must also signal the usage of AI for screening; • otherwise, chatbots achieve more expensive purchase prices than humans.
(Kahraman et al., 2003)	✗	✓	✗	✗	• Use selection criteria of purchasing managers in a Fuzzy analytic hierarchy process.
(Kilincci & Onal, 2011)	✗	✓	✗	✗	• Extensive list of 14 supplier selection criteria. • Develop MS Excel macro for fuzzy AHP.

Table 6.1: Literature Review

	Step being automated				
Reference	Identification	Selection	Negotiation	Spent optimization	Finding
(Shendryk et al., 2019)	✗	✓	✗	✗	• For supplier selection on electronic markets, online mined supplier judgments can be used.
(Lin et al., 2011)	✗	✓	✗	✗	• Considering dependence between selection criteria by combining ANP, TOPSIS and LP.
(Önüt et al., 2009)	✗	✓	✗	✗	• Long-term supplier selection. Considering dependence between selection criteria and linguistic uncertainty in judgment.
(Kellner et al., 2019)	✗	✓	✗	✗	• A visualization of the Pareto-front can be used to reduce the number of manual supplier comparisons to be made.
(Scott et al., 2015)	✗	✓	✗	✗	• Combine supplier selection and optimal order dispatching.
(Guo et al., 2009)	✗	✓	✗	✗	• Machine learn selection criteria from past data.

Table 6.1: Literature Review

	Step being automated				
Reference	Identification	Selection	Negotiation	Spent optimization	Finding
(Wu, 2009)	✗	✓	✗	✗	• Hybrid approach. Machine learn selection criteria from past data and efficiency analysis.
(Chen & Rossi, 2021)	✗	✗	✗	✓	• Stochastic inventory problem with capital constraints.
(J. Shi et al., 2017)	✗	✗	✗	✓	• Purchasing seasonal products with capital constraints.
(Gel & Salman, 2022)	✗	✗	✗	✓	• Bayesian updating for ordering quantity decisions with stochastic provider output.
(Crama et al., 2004)	✗	✗	✓	✗	• Automate supplier negotiation by learning acceptable thresholds for accepting offers.
(Lee & Ou-Yang, 2009)	✗	✗	✓	✗	• Predict the supplier's counteroffer reaction to the purchaser's offer in a selection/negotiation process.

Table 6.1: Literature Review

	Step being automated				
Reference	Identification	Selection	Negotiation	Spent optimization	Finding
(Carbonneau et al., 2011)	✗	✗	✓	✗	• Pairwise prediction of supplier's counteroffer and delivery/return/payment policy.

Overall, these studies highlight successful applications of automation in procurement. However, such studies remain narrow in focus, dealing only with replacing tasks typically done by humans. Surprisingly, the question of determining how much supplier search should be optimally conducted has not been addressed before. This is problematic because, currently, supplier quotations can only be evaluated after an exhaustive examination of the procurement market. Our contribution is therefore directed at a data-driven predictive evaluation of supplier quotations.

6.3 Theoretical Background

6.3.1 Problem Setup

A purchasing manager seeks to purchase K different goods $k \in 1, \dots, K$. The problem is now to find among a set of S_k different suppliers the cheapest offer $p_{s,k}$with $s \in S_k$ that provide the good k. However, for new products, suppliers are unknown and difficult to discover. Getting a quotation from a supplier is time-consuming due to explaining product characteristics and negotiating prices. Hence, the purchase manager must determine how many suppliers S'_k, $S'_k \coloneqq \{s,\ s \in 1, \dots, s^* : i \in S_k\}$, to contact and at which index s^* to stop. This is a multi-objective optimization problem: $\min_{S'_k} [\underset{s \in S'_k}{minp_{s,k}}, |S'_k|]$. We study data-driven approaches supporting purchase managers in determining an optimal stopping point s^*.

6.3.2 Static Stopping Rule: Estimating Reference Price

A simple approach to the above-described problem is the estimation of a reference price $\hat{\mu}_k$, i.e., a preferred buying price, for example, the average market price. This is an approximate version of the ε-constraint method (Miettinen & Mäkelä, 2002) to multi-objective optimization and can be written as $\min_{S'_k} |S'_k|\ s.t.\ p_{s,k} \leq \hat{\mu}_k$. The reference price for new items can be estimated using historic quotation data by linking product characteristics with the average of all quotation data. This linkage can be found using machine learning. Machine learning methods are special cases of optimization problems, which are optimized according to a cost function. Hence, an initial design challenge is quantifying a suitable cost function. To find a suitable cost function, we have chosen to examine the economic consequences of a possibly erroneous forecast. Based on the predicted reference price $\hat{\mu}_k$ and the supplier's offer $p_{s,k}$ the purchasing manager can make three decisions. S/he may, firstly, buy directly, or, secondly, reduce/increase negation efforts, or thirdly, temporarily defer the offer in order to search for a lower quotation from another supplier. We can then assess the decision's impact on various market states,

analyzing the economic consequences of prediction errors. We assume there will always be suppliers that provide prices above and below the reference price. Using this setup, three possible cases of prediction errors ($\hat{\mu} \neq \mu$) exist:

1. $\hat{\mu} \leq \mu$, $p_s \leq \mu$: Some attractive suppliers will be wrongly rejected $WR := \bigcup_k^K \min_s\{p_{s,k}, s \in S_k : p_{s,k} > \hat{\mu}_k\}$. Increases process cost proportional to $|WR|$, lowers purchase cost. If the estimate is too low, no suppliers are discovered.
2. $p_s > \mu$, $\hat{\mu} \leq \mu$: These suppliers are correctly rejected
3. $p_s > \mu$, $\hat{\mu} > \mu$: Some suppliers will be wrongly accepted $WA := \bigcup_k^K \min_s\{p_{s,k}, s \in S_k : p_{s,k} < \hat{\mu}_k \,\&\, p_{s,k} > \mu_k\}$. Decreases process costs, increases purchase cost.

The analysis shows that if purchase costs are an issue, the purchaser should choose a prediction technology that undervalues the market price (case 1.). On the other hand, overestimating the purchase price (case 3.) can increase the purchase but decreases the process cost. Therefore, a purchase manager must determine which performance indicator best balances the competing goals of exploration (finding a better deal) and process efficiency (reducing process costs and higher procurement speed). Section 6.3.3 and Section 6.3.4 discuss a possible solution.

6.3.3 Dynamic Stopping Rule: Without Updating

The primary difference between a static method and a dynamic approach is that the static approach is more likely to inadvertently stop searching even when it is advantageous or not stop searching even when the expected value of the search is low.

To achieve a more targeted resource allocation, the dynamic stopping rule changes the stopping point depending on the probability of sourcing a lower price. The reasoning behind algorithm 1 is quite intuitive. The algorithm computes in line 3) the expected value from searching for lower prices than the current best price. That is, it computes for every possible future p_{s+1} price the probability $\tilde{f}(p_{s+1})$ of obtaining this price from the next supplier. If that price is higher than $p_{best,s}$ the purchaser prefers not to buy; otherwise, the saving is calculated.

Listing 6.1:

set $s = s + 1$ *and iterate from 1. Else stop and choose offer* $p_{best,s}$.

Initialize s=1

1. Obtain p_s
2. Set $p_{best,s} = \min\{p_j | j = 1, \dots, s\}$
3. If $\sum_{p_{s+1}} \left|\min\left(p_{s+1} - p_{best,s}, 0\right)\right| \tilde{f}(p_{s+1}) > c$:

set $s = s + 1$ *and iterate from 1. Else stop and choose offer* $p_{best,s}$.

However, the algorithm does not incorporate new information in its current form.

6.3.4 Dynamic Stopping Rule: with Updating

Now we address the problem of updating the learning algorithm with new data. Updating is important because prediction errors can impair economic outcomes, and Listing 6.1 does not update the recommendations in such cases. This also may not be a good use of available data since, regardless of how accurate the prediction algorithm is, on average, the purchase manager needs to source at least one offer before s/he can make any purchase. The sourced offer could contain valuable information that is otherwise unaccounted for. In addition, individual data series for specific items might be relatively brief, making prediction harder. The static approach's prediction accuracy now hinges on how much predictive power comparable items in the data set provided. On the other hand, the Bayesian approach that we suggest also incorporates new data obtained during supplier search, thus potentially resolving the previously stated issue. In concrete, Bayesian updating allows one to sequentially learn from new quotation data as supplier offers are collected. Our second proposed algorithm uses updating (Listing 6.2):

Listing 6.2:

Initialize s=1

1. Obtain p_s
2. Update $\tilde{f}(p_{s+1} | p_s, p_{s-1}, \dots, p_1)$
3. Set $p_{best,s} = \min\{p_j | j = 1, \dots, s\}$
4. If $\sum_{p_{s+1}} \left|min\left(p_{s+1} - p_{best,s}, 0\right)\right| \tilde{f}(p_{s+1} | p_s, p_{s-1}, \dots, p_1) > c$:

set $s = s + 1$ *and iterate from 1. Else stop and choose offer* $p_{best,s}$.

A particular feature of our approach is that $\tilde{f}(p_{s+1}|p_s, p_{s-1}, \dots, p_1)$ is conditioned on the quotation history at every step, which means that all available information is considered. That also means that an initially deficient estimate could be corrected. The core of our approach is the calculation of the density forecast that incorporates parameter uncertainty using prior knowledge regarding the parameter and is updated sequentially: $\tilde{f}(p_{s+1}|p_s, \dots, p_1) = \int f(p_{s+1}|\theta, p_s, \dots, p_1)\pi(\theta|p_s, \dots, p_1)\, \mathrm{d}\theta$. The so-called posterior can be calculated using the Bayes theorem $\pi(\theta|p_s, \dots, p_1) = \frac{f(p_s, \dots, p_1|\theta)\pi(\theta)}{\int f(p_s, \dots, p_1|\theta)\pi(\theta)}$. All that is needed is a likelihood function $f(p_s, \dots, p_1|\theta)$ and a prior function $\pi(\theta)$. For background on Bayesian methods, see Hoff (2009). Our concrete implementation is described in Section 6.4.4.

6.4 Empirical Application

We evaluated five different approaches to determine the stopping point, namely:

- **Heuristic I**. Only control the process cost by limiting the number of requests for quotation. We set $s^* = 3$, a value typically found at public institutions.
- **Static**. Control the purchase cost by estimating $\hat{\mu}_k$ (see Section 6.3.2). Stop if at the first quote that undercuts the reference price $\hat{\mu}_k$.
- **Dynamic w/o updating**. Calculate the expected gain from searching for a lower price given the current best offer without learning from supplier quotes (see Section 6.3.3).
- **Dynamic with updating.** As w/o updating**,** includes supplier quotes in subsequent calculations of expected gain from searching (see Section 6.3.4).
- **Heuristic II.** Controlling purchase costs by considering many suppliers.

The approaches "heuristic I" and "heuristic II" serve as benchmark cases for controlling process and direct purchase costs.

For the empirical application, we used two data sets. A simulated data set, in which we introduce various kinds of biases in the prediction, to study the robustness of the different approaches. Finally, we employ the algorithm on the real-world data set that motivated our research.

6.4.1 Simulated Data and Scenarios

The simulated data set is generated by randomly drawing μ_k and σ_k^2 from a uniform distribution. Both parameters constitute the true population parameters. We then simulate supplier offers by randomly drawing from a Gamma distribution parametrized with the true parameters. We then compared several scenarios with the prediction technology. For these, we draw the $\hat{\mu}_k \sim$

$Gamma\left(\frac{(\mu_k\tau)^2}{\varepsilon}, \frac{\varepsilon}{\mu_k\tau}\right)$ and $\widehat{\sigma_k^2} \sim Gamma\left(\frac{(\sigma_k^2\tau)^2}{\varepsilon}, \frac{\varepsilon}{\sigma_k^2\tau}\right)$. That means we assume that the prediction technology is of the same quality for both predicted variables. Because of the properties of the Gamma distribution $E(\hat{\mu}_k) = \mu_k\tau$ and $Var(\hat{\mu}_k) = \varepsilon$. The results are for $\widehat{\sigma_k^2}$ analogous. The parameter ε controls the accuracy, or noise, of the prediction technology. The parameter τ controls the systematic direction of bias of the prediction technology. We then specify the following scenarios:

- Low error: ε=0.05, $\tau = 1$
- High error: ε=0.20, $\tau = 1$
- Overestimation: ε=0.05, $\tau = 1.2$
- Underestimation: ε=0.05, $\tau = 0.8$

6.4.2 Real Data Case Study

The case study is from an industrial procurement setting. In concrete, we study procurement of electrical resistors for a large producer of domestic electrical equipment. The data was extracted from suppliers' quotations using text mining. Resistors are inexpensive, costing from a few cents to about 3-5€. Specialized resistors might cost up to €15. Resistors are characterized by different attributes, such as nominal resistance, size, and product quality characteristics. We leverage these attributes to learn the resistor price from its characteristics. The raw data set comprises 201,187 price quotes from suppliers for about 2,400 resistors. Regarding the number of supplier quotations for a specific resistor: the 25th percentile is 18, while the 50th percentile is 53. The study spans the years 2014 through 2019. We improved the comparability of the quotes by adjusting the pricing for 2019. We calculated the average and variance of supplier prices for each resistor type. Using this information, we built two random forests on the training data to forecast each resistor type's average market price and variance. The testing data set includes resistor properties and a collection of offers from numerous vendors. In concrete, we evaluate using 800 unique new resistors.

6.4.3 Evaluation Strategy

For evaluation, we replicate the purchase process. For each resistor $k \in 1, \dots, K$ in the testing data, we predict $\hat{\mu}_k$ and $\widehat{\sigma_k^2}$. This information is utilized to evaluate sequentially each of the S_k offers from simulated and real suppliers. Each approach for determining a stopping point is tested using identical pricing quotations. Therefore, the entire solution space is spanned by a $K \times S$ grid. Each approach is assessed on its ability to efficiently explore the solution space in

terms of achieved purchase costs and procurement process costs. Procurement process costs are approximated by the total number of examined quotes and requests made.

6.4.4 Implementation and Software Used

We now describe the details of how Bayesian updating was implemented. For modeling purchase prices, the *Gamma distribution* is often used (Albright 1977). The Gamma distribution is flexible and can take many forms depending on the parameter values (Hong & Shum, 2006). Hence, in the case of our application, we assume that prices p_{ik} follow a Gamma distribution. In particular, we assume that each type of item, indexed k, has its own price distribution, not necessarily unique, parametrized by s_k and a_k. To model the heterogeneity of prices for different items that may be quite different shaped and scaled, we reparametrize $a_k = \frac{\mu_k^2}{\sigma_k^2}$ and $s_k = \frac{\mu_k}{\sigma_k^2}$. This allows modeling parameter uncertainty in terms of expected value μ_k and variance σ_k^2. We estimate these two parameters for each resistor type. Consistent with the Bayesian paradigm, we assume that the purchase manager can encode prior information about the likely values of the parameters. We define the priors $\pi_1(\mu_k)$ and $\pi_2(\sigma_k^2)$ in such a way that their modes correspond to $\widehat{\mu_k}$ and $\widehat{\sigma_k^2}$. We prefer this specification, as it puts much weight on the initial estimates. The prior on σ_k^2 is assumed to be $Gamma\left(\frac{\widehat{\sigma_k^2}}{\omega-1}, \omega\right)$. We view ω as an additional hyper parameter that governs the weight of the prior. We settled for $\omega = 3$ using a manual search. The prior on μ_k is assumed to follow a PERT distribution (Malcolm et al., 1959). The PERT is a flexible distribution as it is based on a reparametrized Beta model. In addition, the PERT distribution has the advantage that its domain is bounded on the positive scale, in contrast, e.g., to a normal distribution. We prefer PERT for the price distribution because its domain can be bounded on a closed interval. This interval is set to $(0,15]$ according to the typical range of quotes. In principle, other forms of priors are also possible. For example, we could have modeled the prior directly using a Beta distribution. Yet, we settled on the PERT distribution because it can be easily parameterized using only the minimum, maximum, and most likely value. The typical domain of resistor prices defines the minimum and maximum. The most likely value is set to the estimate of the average price $\widehat{\mu_k}$. On the other hand, for the variance, we restrict the domain on values larger than zero and put a higher probability mass on $\widehat{\sigma_k^2}$. Regarding the upper bound on the domain of the prior on σ_k^2, we have more uncertainty. Hence, we chose Gamma distribution as prior for σ_k^2.

All computer code was written in R. For computing the posterior, we used 300×100 Monte Carlo grid approximation for μ and σ^2. The PERT distribution we took from the mc2d package (Pouillot & Delignette-Muller, 2010), machine learning was done with mlr and ranger Lang et al. (Lang et al., 2019), and the future package for parallel computations (Bengtsson, 2021). The stopping threshold c was set to a percentage value of five percent of the estimated product price (relative threshold).

6.5 Results

The results regarding purchase and process costs are depicted in Table 6.1 and Table 6.2. We also tested if the differences between the approaches are significant. For this, we used a paired t-test because all approaches are evaluated on identical simulated/real records and are thus dependent. We find that the average purchase cost for the Bayesian method is significantly higher than for the method w/o (without) updating in both noise scenarios, $t(999) \geq 6.87, p < 0.01$. Also, in the case of overestimation, the Bayesian method is significantly more costly than the method w/o updating, $t(999) \geq 17.55, p \leq 0.01$, whereas in the case of underestimation, the Bayesian method is significantly less costly, $|t(999)| \geq 6.4, p \leq 0.01$. For all the first three scenarios, the Bayesian method yields significantly fewer requests than the distributional method, $|t(999)| \geq 24, p < 0.01$, but for the case of overestimation, the Bayesian method needs more number of requests $t(999) = 15.5, p < 0.01$. Between high error and the underestimation scenarios, there is a significant difference in terms of costs for the Bayesian method, $t(999) \geq 2.46, p < 0.05$. There is no significant difference in costs for the Bayesian method across the remaining simulated scenarios, $t(999) \leq 1.5, p > 0.1$, except that the low error scenario is significantly higher than the high error scenario $|t(999)| \geq 2.33, p < 0.05$.

Table 6.1: Average purchase cost (K_simulated=1,000, K_real_data=800)

Dataset & Scenario	Heuristic I	Static	Dynamic	Dynamic Updating	Heuristic II
Simulated					
Low error	1.957	1.977	1.894	1.894	1.548
High error	1.957	1.962	1.796	1.912	1.548
Underestimate	1.957	1.773	1.621	1.894	1.548
Overestimate	1.957	2.184	1.992	1.904	1.548
Real data					
Random Forest	2.840	2.518	2.520	2.736	2.421

Table 6.2: Average number of requests (K_simulated=1,000, K_real_data=800)

Dataset & Scenario	Heuristic I	Static	Dynamic	Dynamic with updating	Heuristic II
Simulated					
Low error	3 (72%)	2.171 (116%)	3.494 (113%)	3.378	10
High error	3 (98%)	2.950 (140%)	4.193 (108%)	3.250	10
Underestimate	3 (145%)	4.342 (210%)	6.311 (113%)	3.403	10
Overestimate	3 (47%)	1.405 (73%)	2.204 (109%)	3.277	10
Real data					
Random Forest	3 (373%)	11.20 (366%)	10.98 (130%)	3.887	17.186 (573%)

For the number of requests comparing the Bayesian method, there is a statistically significant difference, $|t(999)| \geq 4.22, p < 0.05$, except for high error vs. overestimate, $|t(999)| \leq$

$0.77, p > 0.1$. We also calculated the mean percentage error (MAPE) on all studied settings for reference: For low error 10%, for high error 21%, for underestimate 22%, for overestimate 20%, for random forest 36%. We also tried but did not report other random forests and a NN whose hyperparameters were tuned on a validation set using MAPE, absolute error, and loss functions that penalize for under-/overestimation. However, predictions turned out to be similar.

6.6 Discussion

6.6.1 Results

We found empirically that the static technique has lower purchase costs but higher process costs. The reason is that the static method terminates earlier than heuristic I, hence the purchase price is higher. Process costs are also higher for the real data case, presumably because the random forest underestimated the price average. The simulated results for the underestimating scenario support this. The dynamic rule outperforms the static rule in terms of purchasing costs, not process costs. The dynamic rule without updating has lower procurement costs but slightly higher process costs than heuristic I. So, the dynamic rule keeps searching when there are large expected savings.

Four scenarios of introducing noise and bias in simulation predictions were utilized to assess the dynamic method's ability to correct for forecast errors. The rule with updating reduces process costs in all circumstances except overestimation. In the case of overestimating dynamic without updating is too pessimistic about potential savings, while in the case of underestimation, no updating is too optimistic, similarly to the static rule. The rule with updating is more robust, suggesting that Bayesian updating corrected the initial faulty forecast. That presumably explains why the rule with updating works better in the real data case. The dynamic rule with updating appears to be robust to any prediction bias in the simulated data for purchase costs, as indicated by the non-significant t-tests. This observation suggests that the direction of bias is unimportant for the dynamic approach with updating, although it appears essential for the static and dynamic rule without updating. We find it expected that overall differences between the scenarios for the Bayesian method are non-significant for purchase costs but significantly different in terms of process costs. It shows that the Bayesian method is robust towards deficient predictions that enter as an argument; such deficient forecasts are then corrected by exploring more supplier offers. The method w/o updating has lower purchase costs in the case of underestimation, although this comes at higher process costs. That finding implies that the dynamic stopping rule w/o updating is not recommendable. In the simulated scenario, the distribution of received quotes belongs precisely to the same family of statistical

distributions used to calculate the dynamic stopping rule. In contrast, in the real data application, we used the Gamma distribution to approximate the real distribution of prices. Despite being an approximation, our approach also extends to the real data case. Estimating distribution parameters using machine learning works, despite low predictive accuracy, as indicated by the high MAPE for the random forest. Nevertheless, the dynamic stopping rule with updating benefits from information included in obtained supplier quotes.

6.6.2 Limitations

Our method applies to many procurement situations but is based on explicit assumptions: a) Obtaining a new request for quotation is costly, and b) an offer can be deferred at no additional cost. In concrete, a) is plausible because of all the search-related costs incurred from scouring the market for the best alternative (Choudhury et al., 1998). Assumption b) requires supplier quotations to be valid for a certain period (e.g., if suppliers submit a binding price quotation). This may not apply to some types of products: e.g., seasonal products, temporary discounts, commodities. A workaround is to gradually increase the termination criterion to reflect the effect of delaying.

In sum, these assumptions put weak limits on the applicability of our approach. Even so, some reservations should be made. The supplier's strategic behavior is currently being disregarded (e.g., concerning supplied price offers). This study assumed that the supplier's final best offer is the decision input, ignoring any bargaining premium. However, in practice, a purchase manager should consider negotiation strategies Oliver (Oliver 1996). We did not model price changes, which are essential for real-time spot market purchases (e.g., energy), but can be neglected if prices are temporarily stable. Also, purchasing for immediate production needs limits the ability to delay purchases. Purchasing can also be subjected to additional objectives when considering supplier properties (e.g., lead times, quality). Scalarizing (Miettinen & Mäkelä, 2002) and constructing a joint probability function of these properties may be a way to address this issue. Finally, we did not investigate purchase costs (delivery, logistic, and storage costs) as they are conceptually different from the general procurement process.

6.6.3 Implications for Practice and Research

The findings are significant for purchasing managers since both the w/o updating and the Bayesian method offer several advantages. First, these techniques can be used to increase average procurement speed while also reducing average costs. As a result, the strategy keeps control over both purchase and process costs. Second, the techniques justify prioritizing specific procurement projects. In concrete, it provides managers with a tool for communicating when

procurement efforts should be expanded or when they can be halted or reallocated between projects depending on the expected value of further searches. Third, procurement managers can make more precise statements about the value their department contributes to the organization's bottom line using the proposed technique. Finally, the approach can also be used to track and direct the efforts of the procurement department and the efforts of individual staff members. Purchasing managers can use our algorithms as a self-service-analytics solution (SSA) within standard procurement software solutions (Allal-Chérif et al., 2021). Future research could focus on optimally incorporating our proposed solution in an SSA concerning socio-technic design characteristics. For instance, it is unknown whether purchasing managers view the algorithmic solution positively or whether they would follow the algorithmic recommendations at all.

7. A Multi-Objective Particle Swarm Optimization Framework for Operations Management

Julian Sengewald (TU Dortmund)
Richard Lackes (TU Dortmund)

Abstract. There is ongoing research on the problem of how to best combine predictive modeling and optimization. This is especially important in operations management, where there are complex business processes to be optimized. We propose a framework based on evolutionary computing with multi-objective particle swarm optimization and on the design of the fitness function according to the business operations to be optimized. By doing so, one can optimize a range of interesting problems using neural networks that would be otherwise hard to handle with classically supervised learning.

7.1 Introduction

7.1.1 Predictive Machine Learning and Neural Networks

The use of machine learning (ML) in operations management is growing (Vanderschueren et al., 2022). For example, ML helps fast-food and retail organizations adjust staffing levels based on expected customer demand. This avoids overstaffing and understaffing. In predictive maintenance, companies use ML to predict potential component failures and ensure timely repairs. This can result in significant cost savings. All these applications share the common characteristic of predicting quantities before they occur, allowing the organization to take appropriate action in a timely manner. This approach addresses operations management issues through the "predict first, optimize later" paradigm, using various machine learning methods (Vanderschueren et al., 2022). The *"predict first, then optimize"-paradigm* bases optimization on predictions, obtained from analyzing data for predictive patterns between input features and the outcome using ML. A common ML technique are deep neural networks (DNN) (Kraus et al., 2020). DNN have neurons as their basic units. These neurons are arranged in layers and connected through links. Each link has weights $w_{i,j}$ connecting layers i and j. The first layer of a network carries the input data. The last layer is the output layer, representing the prediction of the DNN. These two layers are connected via one or more hidden layers. Each layer consists

of a fixed number of neurons m_i in layer i. The hidden layers process the output from their preceding layer, wherein each neuron receives $n_j = \text{act}\left(\sum_{i=1}^{m_i} \cdot w_{i,j} n_i\right)$ as input. Here, $act(.)$ is a non-linear function referred to as the activation function. Training a DNN requires solving an optimization problem by adjusting the weights $w_{i,j}$ such that the network's prediction matches the observed data.

7.1.2 Supervised Learning and Its Limitations for Operations Management Problems

Supervised learning algorithms, such as DNN, have limitations when it comes to solving operations management problems (Elmachtoub & Grigas 2022). For example, the current use of ML and DNN provides solely predictive information and lacks the full capability of managerial decision-making (Kraus et al., 2020). Organizations often have different downstream implications arising from predictions, rendering a mismatch between the loss function used in training and the true optimization objective of the business operations (Elmachtoub & Grigas 2022). Many supervised prediction applications in operations management omit to include the uncertainty associated with decision-making (Kraus et al., 2020).

Gradient descent is a common method for training DNNs (Kraus et al., 2020); its success in solving an ML problem largely depends on the availability of labeled training data and whether the loss function used accurately reflects the intended optimization objective. In contrast, in the context of operations management problems, loss functions typically fail to sufficiently capture the true optimization objective mathematically. For instance, in regression tasks, a common choice for the loss function is mean squared error $L(y, \hat{y}) = \frac{1}{2N}\sum_i^N (y_i - \hat{y}_i)^2$. Optimizing for this loss function helps the DNN get closer to the conditional expectation of the target y_{target} given the input, since errors that are equally far above or below the prediction are given the same amount of weight. Yet, the implications of prediction errors for the interlinked decision problem are not necessarily symmetric. We call this *complication 1*. In personalized pricing, for example, an overestimation of the reservation price is worse than a underestimation, because the latter leaves some margin, at a lower price, for the business. Hence, the quadratic loss function wrongly reflects the preferences of the business owner.

An approach to overcome these limitations is to design asymmetric losses. For example, the loss function could be adjusted so that it penalizes positive prediction errors more than negative prediction errors: $0.5\left(y_{\text{target},i} - y_{\text{prediction},i}\right)^2$ if $y_{\text{target},i} > y_{\text{prediction},i}$ and $0.5\left(y_{\text{target},i} -$

$y_{\text{prediction},i}\big)^2 \cdot 2$ if $y_{\text{target},i} < y_{\text{prediction},i}$. This loss function penalizes positive prediction errors twice as much as negative prediction errors. Yet, there are a few more factors to consider in practice: i) the weights that were attached to the loss need to be known; ii) the same is true for the actual label, the reservation price, such that the prediction error can be calculated. While both may be addressed by using some suitable proxies that can be used instead of the principal quantities (Vanderschueren et al., 2022), the overall quality of the solution depends on how well using the proxy approximates the overall problem when the principal quantities would have been available. We call this *complication 2*, the proxy approximation challenge. Instead of using proxies, if they exist at all, a more straightforward approach would be to simulate business operations and optimize an ML model-based on the repercussions of its prediction. This method does not require knowing the weights attached to the loss or the actual label.

Training a DNN with a simulation of business operations comes with its intricacies, especially when feeding the simulation output into the network. One way to address this challenge is to use a standard gradient-based predictive model and adjust the network architecture according to the simulation performance. In other words, the simulation is only used to tune the hyperparameters of the network. However, this method has drawbacks. First, it may result in suboptimal performance and inaccurate predictions since the network is not directly optimized for the actual objective at the level of the weights. Second, it may not be suitable for multi-objective problems where it is difficult to assign weights to each objective, linearize the problem, or apply the ϵ-constraint method. Therefore, not all objectives can be accurately optimized.

Overall, we conclude that there are four common edge cases for supervised learning algorithms where the prediction task is either not well-defined, not well-aligned with the business objective, or not well-measurable (see Table 7.1). These edge cases pose challenges for designing supervised learning models. As we have discussed above, there are many compromises to take to make supervised learning feasible in such circumstances; see also Table 7.1.

Table 7.1: Edge Cases for Supervised Learning Algorithms

Problem	Description	Examples
Complication 1	No natural quantity to predict	no label; sequential problems
Complication 2	Prediction errors are costly and imprecisely measurable	time lags, guesstimates, unreliable data
Complication 3	Ill-defined loss function	complex relationship between ML prediction and business outcome
Complication 4	Multiple objectives to optimize	multi-objective problem, trade-offs

Accordingly, we suggest that training NNs with gradient descent and standard loss functions may not optimally solve operations management problems involving complex consequences of prediction errors. In the next section, we explore alternative optimization methods that could overcome these limitations.

7.1.3 Case Study

We discuss the application of our optimization framework in the context of supporting procurement processes. The case study is based on the problem of sourcing suppliers and procuring items from them. The problem is that this overall process should be cost-efficient. We based our study on a dataset from the author's previous research (Sengewald & Lackes, 2022).

The cost of total procurement includes the direct purchase cost of the item and the procurement process costs of sourcing suppliers. Supplier-sourcing is costly in an industrial environment because it involves the working time of procurement staff in locating and contacting suppliers, as well as activities such as evaluating bids and the technical qualifications of items. All this processing can add up to a significant amount. When sourcing suppliers for many related items regularly, supporting the buyer with an intelligent decision support system can make the whole process more cost-effective. This is because information can be used to guide the sourcing process. For example, with the help of such an intelligent system, the agent can decide whether it is worthwhile to conduct further research and obtain additional quotes (lower purchasing costs) or not. One difficulty with process costs is that they may be difficult to quantify, and the buyer may wish to minimize the total number of suppliers contacted to avoid contacting too many suppliers unnecessarily (optimization problem). Therefore, we investigate how purchases can optimize their decisions when purchasing items from different suppliers. Overall, this case

study presents all the features of the complications referred to before (see Table 1). There is no natural quantity to predict because the purchasing agent needs to find a low-priced supplier while balancing the purchase costs (Complication 1). One may use the average price of quotations for similar items as a rough guiding point. However, using the average as a stopping price may not provide the best approach to the underlying optimization problem. Therefore, we propose a method that models the consequences of the prediction so that one does not need to specify a label. Predicting an above-average market price could lead to the purchasing agent wrongly rejecting over-average offers, while underestimating may lead to wrongly accepting over-average offers. Which error is preferable is only decidable once the prediction errors are quantifiable (Complication 2). Yet, these prediction errors are not directly linked to the actual optimization objective, which consists of trading off low purchases with overall process costs, which cannot easily be represented by a standard loss function (Complication 3) and is multi-objective (Complication 4).

7.1.4 Related Literature

As our case study is based on procurement, we briefly review the status of ML applied to procurement and purchasing. We then summarize the relevant literature on ML and operations management.

Purchasing has increasingly turned to the use of AI (Allal-Chérif et al., 2021). Examples of potential uses of AI in purchasing include automation of repetitive tasks, process monitoring, supplier selection through matching systems, and decision support (Allal-Chérif et al., 2021). Prior review research combines academic literature and industry surveys to explore procurement analytics solutions from both the supply (vendor) and demand (user) sides (Handfield et al., 2019). By contrasting existing solutions from academic literature and industry with the needs identified by procurement professionals, the research aims to identify gaps and opportunities for improvement in current procurement analytics practices. This effort is particularly relevant when examining the sourcing phase, which, according to prior literature, consists of five steps: 1. specification definition; 2. supplier identification; 3. requests for quotation; 4. negotiation and selection; and 5. contracting (Spreitzenbarth et al., 2024). Each of these phases has received varying levels of support from sourcing analytics. For example, while there are several solutions for supplier selection and negotiation, both in the academic literature and in industry software (Allal-Chérif et al., 2021), there is a lack of tools to support purchases in the activities preceding these steps. Some tools support the process of supplier discovery, also known as supplier scouting, by focusing on advanced AI-based search (e.g., incorporating

unstructured data) (Guida et al., 2023). The ability of these tools to control search costs and gather critical supplier data supports their effectiveness. These tools can help streamline the supplier-sourcing process. In addition, analytical support for supplier pre-qualification aids in tactical decisions (Guida et al., 2023). In contrast, practitioners highlight the critical role of price benchmarking and the value of improved market intelligence as valuable insights (Handfield et al., 2019). Overall, the literature emphasizes the need for solutions that support purchasers at the operational level (Guida et al., 2023) and provide procurement intelligence (Handfield et al., 2019).

ML for Operations Management. First, data in operations management typically has many categorical features, which can pose challenges for gradient descent optimization used in DNN due to discontinuities when using one-hot encoding (Kraus et al., 2020). Custom pre-processing via embeddings provides a better solution for handling categorical features, which are relevant for many applications in operations management (Kraus et al., 2020). Another line of literature revolves around the predict-then-optimize framework. In operations management, decision-making should account for outcome uncertainty due to outcome variability. However, many supervised ML methods discard this information for decision-making by predicting only the conditional expectation (Bertsimas & Kallus, 2020). An approach that makes better use of the available data is to use a weighted expectation of the cost of a decision, where the weights are learned from relevant covariates associated with the outcome (Bertsimas & Kallus, 2020). Alternatively, if the objective function is linear, there exists a surrogate loss function that can be used to incorporate decision error into training (Elmachtoub & Grigas, 2022). Another strategy is to construct weighted versions of conventional loss functions (e.g., weighted cross-entropy) that are differentiable and thus can be used by conventional ML methods (Vanderschueren et al., 2022).

7.1.5 Summary of Contribution

In summary, supervised learning via DNNs can be challenging in certain edge cases (see Table 7.1). Overall, resorting to standard gradient-based optimization for DNNs may lead to many compromises when using a standard DNN training procedure. In the following sections, we describe that these compromises are not necessary with a gradient-free optimization method: Particle Swarm Optimization (PSO). We also show that it is straightforward to optimize for multiple objectives using the PSO method. By doing so, we extend previous approaches towards combining prediction and optimization simultaneously for the actual optimization objective 1) with an additional procedure that does not need an explicit label and 2) is also able

to optimize multiple objectives. We also explain when this is the case, such as when there is no natural label to predict, and one seeks to optimize many objectives. Finally, we illustrate our approach using a case study in procurement, which draws from previous research (Sengewald & Lackes, 2022).

7.2 Description of Framework: Concepts and Measures in Multi-objective Optimization

7.2.1 Multi-objective Optimization

Balancing Cost and Efficiency in Supplier Search. Multi-objective Optimization (MOO) involves optimizing multiple objectives simultaneously. For example, one might want to optimize purchase costs and procurement process costs simultaneously. When evaluating suppliers, minimizing direct purchase costs requires assessing a large pool, while minimizing procurement process costs involves keeping the supplier pool small (e.g., less effort on supplier qualification). Formally, this problem can be represented as $\text{argmin}_s = (Q(s), s)$ where $Q(s) = \min\{p_1, \dots, p_i, \dots, p_s\}$, and s is the number of suppliers contacted, with p_i being the corresponding price offer from the sth supplier. In MOO, considering multiple dimensions is necessary, unlike in single-objective optimization where sorting by objective value suffices for ranking. For this, the concept of Pareto optimality is used. Pareto-optimal solutions are those in which no other solution dominates them.

Pareto-Optimal. Formally, a solution x is Pareto-optimal if there is no other solution x' such that x' is better than x in all objectives and strictly better in at least one objective. Because an increase in one objective value can make up for a decrease in another, there are multiple optimal solutions available to the decision-maker. The collection of all such solutions that are not subject to dominance by any other solution is known as the Pareto-optimal front.

Pareto-Front. The set of all Pareto-optimal solutions is called the Pareto-front. Formally, the Pareto-front is the set of all solutions that cannot be improved on one objective without worsening at least one other objective. All points on the Pareto-front are equally good; yet several such Pareto-fronts can be computed by an algorithm depending on how they were parametrized. The hypervolume indicator, for instance, indicates that when ranking two Pareto-fronts produced by various algorithm configurations of the architecture of a DNN, the Pareto-front with a higher volume has a better ranking because it contains solutions that dominate another Pareto-front.

Hypervolume indicator. The hypervolume indicator is a measure to compare Pareto-fronts. It is the volume of the area where the Pareto-front dominates, and the volume is calculated relative to a pessimistic reference point (Zitzler et al., 2003).

7.2.2 Gradient-free Multi-objective Optimization with Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a heuristic optimization technique. In PSO, a population of potential solutions (called particles) moves around the search space, and each particle's movement depends on both its own trajectory and the trajectory of the swarm. The position of the best solution so far and the best position in the particle's trajectory both affect the trajectory of the particle (candidate solution). Approaching the best solution found is analogous to gradient descent: pointing in the direction of the optimum. However, instead of the gradient, a swarm is used to explore the location of the optimum in the solution space. The role of the swarm itself is both to explore new solution areas and to exchange information with other particles, allowing them to converge on the optimum solution.

There are several extensions of the single-objective algorithm that can extend PSO to the multi-objective case (Coello & Lechuga, 2002). First, rather than a multidimensional gradient, the particles will naturally gravitate toward the Pareto-optimal set, which is the best solution so far. Since there are several such solutions, the challenge is to determine which one to choose to determine the direction in which an individual particle must move. For determining the global best solution, a leader selection mechanism must be ap-plied (Kruse et al., 2022). This can be done using sigma scaling, which finds the particle in the Pareto-optimal set that is closest to the particle in terms of the outcome space (Mostaghim & Teich, 2003). Sigma scaling also moves the particles closer to the Pareo-front (Mostaghim & Teich, 2003). The formula for sigma scaling:

$$\sigma = \frac{o_1^2 - o_2^2}{o_1^2 + o_2^2}$$

, where $o1$ and $o2$ are the outcomes of the two objectives. The particle with the highest σ value is the one closest to the particle in the outcome space (Mostaghim & Teich, 2003).

An archive stores the solutions generated by the continually evolving swarm, which are then utilized to construct the Pareto-front (Kruse et al., 2022). During each iteration, a comparison is made between all the solutions in the archive and the current set of solutions. If a new solution

dominates a solution in the archive, a particle from the archive is updated with that solution from the current swarm (Kruse et al., 2022).

7.3 Application to Case Study

7.3.1 Description of Methodology

To evaluate our approach, we apply our methodology to a real procurement dataset.

Features and data used. The data source is the same as in previous research by the authors (Sengewald & Lackes, 2022). The data is a collection of technical items, electrical resistors, which can be described according to a technical standard but are produced and sold by different suppliers. The data contains thus the technical features, which are common among all resistors. The resistor takes different specifications over the common attributes, such as resistance, wattage, or form factor which allows engineering features that can be used to predict quantities for new resistors (see description of decision modeling below).

The data also contains the price of the resistors from different suppliers. The suppliers themselves are not characterized further. The data is split into a training (n=400) and a testing set (n=200). Train and testing data set have no overlap in terms of the part number of the item. The testing dataset thus comprises the scenario of procuring completely new items from (unknown) suppliers. The features have undergone the same processing as in (Sengewald & Lackes, 2022). For usage in the DNN, the features are normalized to the interval [0,1] by dividing by the maximum value of the feature. In total, after pre-processing categorical features by one-hot-encoding, the data contains 20 features. The features are the resistance, wattage, tolerance, temperature coefficient, form factor (area and volume), material (4 types), resistance, wattage, tolerance (3 levels) and additional five derived by relating wattage and resistance to the size and volume of the resistor. To each resistor that has been PartNumber belongs a set of prices from different suppliers, which have been adjusted for inflation for better comparability because the source data stems from different years.

We use a *deep neural network* with the technical specifications of each part number as input features. This input is propagated through the DNN in a forward pass. In contrast to conventional DNNs, a PSO algorithm adjusts the weights.

Decision modeling. The DNN predicts a stopping price that if undercut leads to the stopping of further searches for suppliers. After a stoppage has occurred, the lowest-priced supplier so far will be chosen to procure from. This price thus determines the purchase costs. Naturally, the

lowest purchase price will be the offer that led to the stopping, but not the predicted price itself. Also, the number of searches performed will determine the process costs.

The tuning of the DNN consisted, firstly, of tuning the architecture of the network (size of hidden layers). Secondly, different parameter values were also tested for the underlying MO-PSO algorithm, as its optimization capacities are the primary research question of this paper.

Objective/fitness function. The *fitness function* simulates the repercussions of DNN prediction on operational outcomes. The DNN produces in the forward pass a stopping price for each part number p_S. Using this stopping price, the quotation prices are sampled randomly and sequentially from the database of historical prices. Then, if the obtained price is lower or equal to the stopping price, the process stops. Then the purchase costs are the lowest price of all offers obtained so far before the stopping price was reached. Necessarily, the purchase price is the last price obtained. The process costs reflect the stopping index, which is, the number of offers that were seen until the stopping price was reached. Both quantities, purchase and process costs, yield a tuple $(\text{PartNumber: } p_{\text{purchase}},\ p_{\text{process}})$. This tuple represents the quality of a particular solution. The above calculations are repeated ten times for each PartNumber and averaged to account for randomness in the ordering of price quotations. By doing so, we obtain the expected performance of the stopping process obtained from the DNN. The overall algorithm that calculates the fitness of a solution is depicted in Listing 7.1.

Network architecture and training. We used a DNN consisting of two hidden layers and one output layer. The number of neurons in the hidden layers is a hyperparameter of the model. We tested $m_1 \in (18, 8)$ and $m_2 \in (8, 5)$ for the number of neurons in the first and second layer. The output layer consists of a single neuron. The activation function of the hidden layers is the leaky rectified linear unit (leakyReLU) and the output layer is linear. The output layer thus predicts a continuous value. The training procedure of the DNN was governed by the MO-PSO. Here we used for the cognitive parameter $c_1 \in \{1,\ 2,\ 3\}$ and for the social parameter c the same values as for the cognitive. Setting the social and cognitive parameters equal provides the optimizer with an equal opportunity for exploration and exploitation. For the inertia, we used $w \in \{0.9, 0.7\}$, and for the number of iterations $maxits = \{50, 250, 500\}$. We used a random grid search on the grid spanning these hyperparameters using 16 evaluations, which was the number of evaluations performed after running the experiment for 10 hours.

Weight initialization We use "Glorot"-initializer for the weights of the DNN (Glorot and Bengio, 2010). In this initialization, the weights are drawn from a normal distribution with mean 0 and variance $1/n$, where n is the sum of number of neurons in the previous and subsequent

layer (Glorot and Bengio, 2010). These weights represent the initial solution for the DNN. By randomly sampling the weights according to the initializer, each individual in the swarm represents a different initial solution for the weights of different DNNs. Note that each individual represents the same architecture of the DNN, but different weights that are optimized by the MO-PSO algorithm.

The forward pass of the DNN must also be self-implemented. A separate backward pass is not necessary as it is replaced by the MO-PSO algorithm. The MO-PSO algorithm was self-implemented in R using the methodology described in the literature Kruse et al. (2022). An external R package, EOF, was used to compute the hypervolume. The EMOA package was used to compute Pareto dominance.

Listing 7.1: Evaluation procedure.

---

1. For each partNumber ∈ E do
2. $p_{\text{stop}} \leftarrow$ get stopping price from DNN for f
3. For rep ∈ $1,\ldots,10$ do
4. $q_1 \leftarrow$ obtain the first offer randomly
5. $s \leftarrow 1$
6. $Q \leftarrow q_1 \cup Q$
7. $bestOffer \leftarrow min(Q)$
8. While $bestOffer \geq p_{\text{stop}}$ do
9. $q_s \leftarrow$ get next offer
10. $Q \leftarrow q_s \cup Q$
11. $bestOffer \leftarrow min(Q)$
12. $s \leftarrow s + 1$
13. End While
14. Append named tuple $(partNumber, rep{:}\, s, bestOffer)$ to temp results
15.
16. $(partNumber, bestOffer) \leftarrow$ average temp results over reps 1,…,5
17. Append named tuple $(partNumber, bestOffer)$ to result.
18. End For
19. Return(result)

7.3.2 Results

First, we compute performance metrics for each model. Each model is evaluated based on the training and test sets. The model differs depending on the architecture of the DNN (i.e., the number of neurons in the hidden layer) and the hyperparameters of the MO-PSO learning procedure (i.e., $c1$, $c2$, and w). The performance metrics are the RMSE and the hypervolume of the Pareo-front. The hypervolume is computed using the eof R package.

In Table 7.2, we show the top 4 models by hypervolume on the training set. Sorting by hypervolume on the train set corresponds to the usual procedure of hyperparameter tuning and

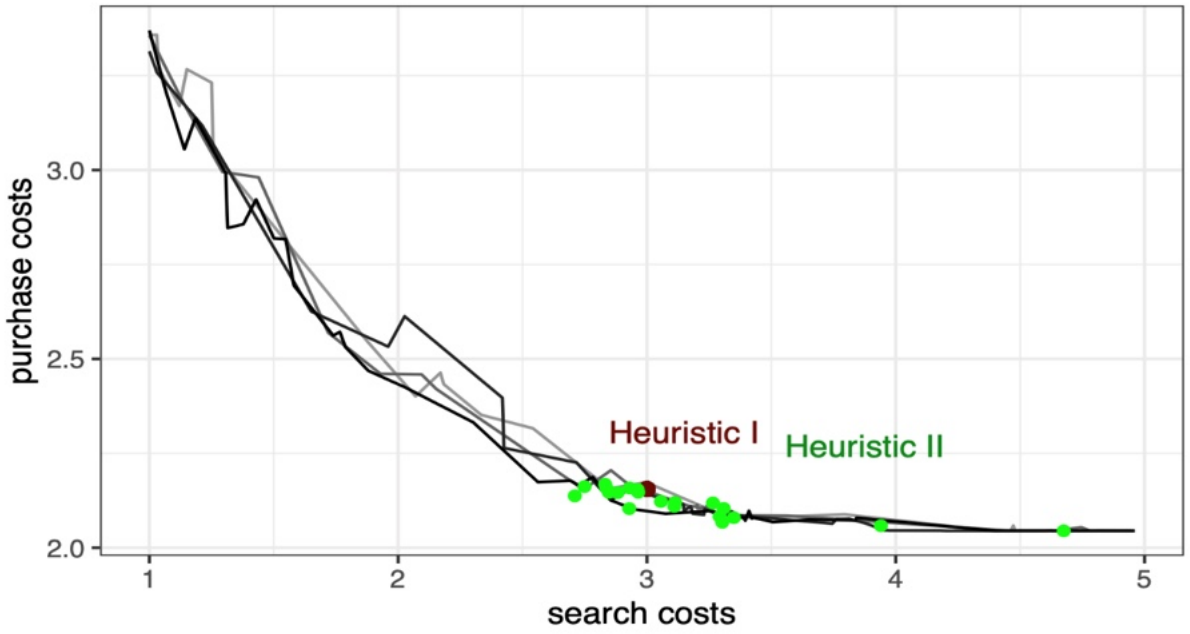
picking the best model, which is then evaluated on the test set. Using hypervolume as a metric for model selection allows for MOO of the model selection process.

Table 7.2: Top 4 models by hypervolume on the training set. The hypervolume is computed using the eof R package. The RMSE is computed on the training and test sets. The architecture of the DNN is given as number of neurons in the hidden layers and output layer.

Parameters					**Hypervolume**		**RMSE**		
c1	c2	W	Its	architecture	Train	Test	Train	Test	#solutions
2	2	1	50	18-5-1	349	340	6	7	59
1	1	1	250	8-8-1	349	270	20	21	100
2	2	1	500	8-5-1	349	270	318	318	69
2	2	1	50	18-8-1	349	340	3	3	56
Keras				8-5-1			1	2	

Since the hypervolume indicator is a multidimensional metric characterizing a whole solution space, we also plot the Pareo-front, which visually depicts each solution in the archive. The Pareo-front is plotted in the search and purchase cost spaces (see Figure 7.1). The search cost is the number of episodes the agent has to search for a product before purchasing it. The purchase cost is the price of the product. Both are averaged over different items in the testing set. The Pareo-front is plotted for the top 3 models by hypervolume on the training set and their resulting performance on the testing set. The result for the static search strategy is also plotted. The static search strategy is a simple heuristic that always searches for the product in the first three episodes and purchases the cheapest found after three quotations have been seen. The static search strategy is used as a benchmark for the performance of the MO-PSO models Performances (Algorithm 1) on testing set. Pareto-fronts for top 3 models in terms of hypervolume during training. Heuristic I is a static search strategy (search = 3). Heuristic II label is $y_i = \frac{1}{S}\sum_{s=1}^{S} p_{i,s}$ trained by MSE loss.

Figure 7.1: Pareto-fronts for top 3 models in terms of hypervolume during training. Heuristic I is a static search strategy (search = 3). Heuristic II label is $y_i = \frac{1}{S}\sum_{s=1}^{S} p_{i,s}$ *trained by MSE loss.*



In Figure 7.1, we compare the MP-PSO algorithm with two other benchmark strategies, Heuristic I and Heuristic II. The static search strategy is a simple heuristic (Heuristic I) which always searches for three episodes and purchases the cheapest offer found after three quotations have been seen. Overall, we see in Figure 1 that, using our approach, we can find solutions that are better than the Heuristic I strategy (marked in red). We also see that there are different Pareto-fronts produced by differently configured models, as these Pareto-fronts are shifted in the search and purchase cost space. Finally, we see solutions produced by the Keras model (green) that are clustered around Heuristic I. We also see that the solutions produced by Keras using MSE loss and average price as labels are not diverse, as they barely vary in their objective values. Overall, low diversity is expected, as each of the 20 models produced by Keras is trained to predict the same label. We apply the Keras model to the test data, so some variation in the predicted labels is expected, and therefore, these different stopping prices also lead to different search costs.

7.4 Discussion

The use of machine learning in organizational operations is growing in popularity. Predictive analytics used in operations management often follows a "predict first, then optimize"-pattern. This means that businesses use past data to construct a predictive ML model. This approach

works well when there is a specific quantity, such as demand or component failures, that can be accurately predicted. By accurately predicting this quantity, businesses can efficiently plan and react to operational needs. However, not all problems fit neatly into this pattern, and it may not be clear what to optimize if there is no naturally linked quantity to the operational outcome.

In this paper, a framework based on evolutionary computing is proposed that can be used for problems that do not fit into the above "predict first, then optimize"-pattern. Similarly, as with supervised learning, this approach utilizes historical data and develops an ML solution. The framework involves training a DNN based on simulating the effects of its predictions. We illustrate this framework by applying it to the domain of procurement, where organizations aim to optimize for conflicting objectives: purchase costs and process costs. To achieve this, we train a DNN to predict stopping prices using multi-objective particle swarm optimization (MO-PSO). This means that instead of following the traditional "predict first, then optimize" paradigm, we can directly optimize for the desired outcome. This approach also allows for the handling of MOO problems. Furthermore, an explicit label is unnecessary, and it suffices to model the decision problem. By modeling the actual business problem as the target of the training procedure, we may also achieve better results after the model has been deployed. This approach not only increases the transparency of the DNN model but also allows the system's logic to be transparent to the user, as the system minimizes errors associated with its decision.

As with every research, there are limitations. First, the framework was only tested with one case study, and in other optimization problems, the results may be different. Second, it is not said that the presented approach cannot be improved by further research, e.g., turning the learning problem into a sequential one where past price quotations are also used for learning.

In sum, the proposed framework offers several advantages. First, it allows for the handling of MOO problems (e.g., purchase costs and process costs). This is difficult to achieve with conventional supervised machine learning algorithms. Second, the approach eliminates the need for an explicit label and models the decision problem directly. This not only simplifies the training procedure but also increases the alignment of the predictions to the actual optimization problem and transparency of the DNN model. Lastly, by modeling the actual business problem as the target of the training procedure, we may achieve better results after the model has been deployed. This is because the model incorporates the entire decision-making process into the training procedure, allowing for more accurate and optimized predictions.

8. Robo-Advisory and Algorithmic Trading via Evolutionary Discretization and Rule-Mining

Sengewald, Julian and Lackes, Richard, “Robo-Advisory and Algorithmic Trading via Evolutionary Discretization and Rule-Mining” (2024). AMCIS 2024 Proceedings. 12.

9. Bike-Sharing Station Placement: Spatial Analysis and Data Mining of Network Design Characteristics

Sengewald, Julian and Lackes, Richard, “Bike-Sharing Station Placement: Spatial Analysis and Data Mining of Network Design Characteristics” (2024). AMCIS 2024 Proceedings. 10.

C: Conclusion

“The first principle is that you must not fool yourself, and you are the easiest person to fool."

Richard Feynman

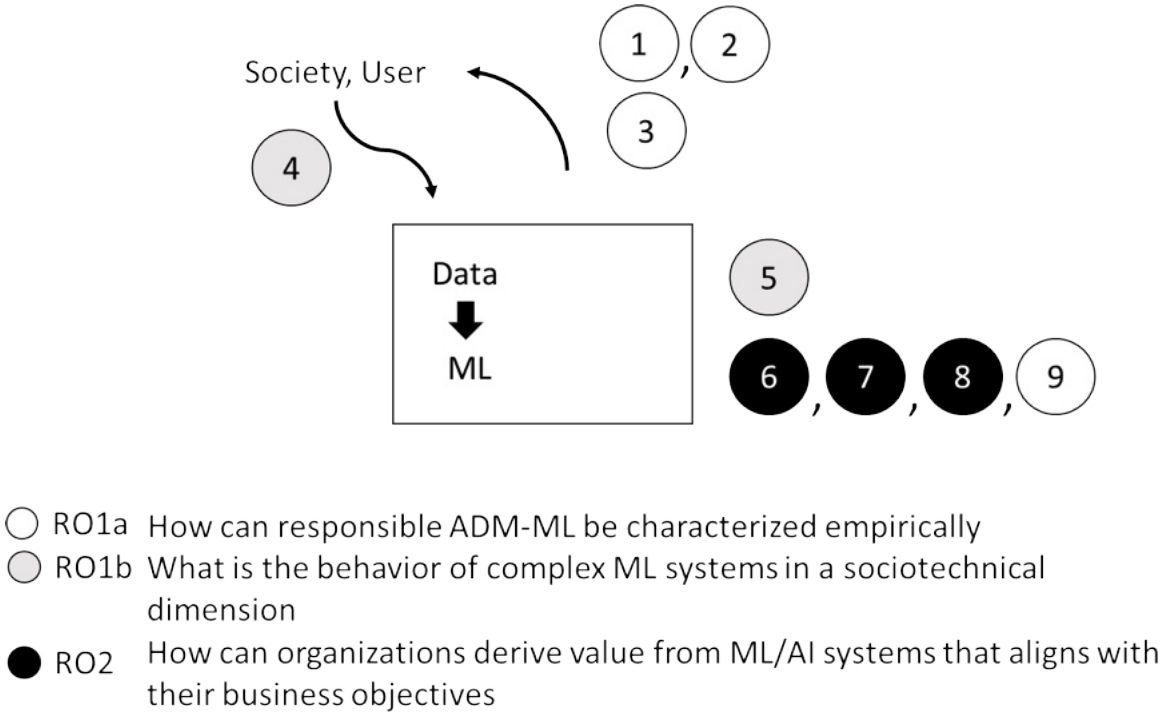
10. Discussion

10.1 General Discussion

Overall, this thesis has addressed ADM-ML in organizations from a dual viewpoint: the instrumental (value-oriented) and humanistic (responsible, ethical) objective of ADM-ML usage (Sarker et al., 2019). From an organizational perspective, both objectives must be considered together because focusing solely on one – particularly the instrumental – without advancing the humanistic objective is unsustainable (Ghasemaghaei & Kordzadeh, 2025; Nima Kordzadeh & Ghasemaghaei, 2021). Therefore, this thesis takes a holistic approach by exploring the role of AI/ML in decision-making, focusing on the multiple relationships between technical and social components in complex AI/ML systems.

In Figure 10.1, paths are depicted stylizing the relationship between the contributions in the thesis and the technical components (data and ML) and the user/society that interact with the technical components. Contributions C1-C3 study the path of how AI/ML-based systems exert influence over humans (AI/ML → humans). C4, although also having components where the AI/ML exert agency over humans, the primary difference to C1-C3 is that the affected humans of AI/ML systems also have agency over the machine that makes the decision affecting them (humans → AI/ML). Finally, C5 studies technical components of how data affects AI/ML (data → AI/ML) as data is an essential input to ML. Contributions C6-C9 study ML-ADM artifacts that are embedded in the contextual problem, designed for specific organizational problems. All contributions address research questions and the alignment of each contribution to research question (RQ) is also depicted in Figure 10.1.

Figure 10.1: Research questions and conceptual relation of contributions.



RO1a addresses the question how responsible ADM-ML can be characterized empirically. This means making recommendations based on behavioral research. With this regard C1 makes important normative recommendations that can be used by other researchers to design experiments that study fairness preferences in ADM. Firstly, we found that the set of fairness metrics may not be exhaustive in the body of researched literature. In addition, scenarios are sensitive to many factors. Therefore, there may be a need for a scenario bank containing calibrated and parametrized scenarios. Another factor is the costs of wrong decisions. The taxonomy of gateway and selection decisions may be helpful in this regard. The primary difference between these two decision types is that gateway decisions incur cost if a wrong decision is made as part of the procedure. Selection decisions have a fairness calculus consideration through having an inherent characteristic that not all qualified candidates can be selected. Thus, mostly the wrong selection of unqualified candidates is a concern in the fairness calculus if this occurs at the expense of qualified candidates. This hypothesized relationship was tested in an empirical experiment considering the case of selection decisions in recruitment, where prior literature had mostly focused on gateway decisions. The results in C2 suggest that participants perceived the equalized odds definitions (FPP and TPP) as fairer than the other definition. Furthermore, we found that DP was the second most preferred fairness metric. However, these results contradict those of Srivastava et al. (2019) who found that DP was considered the fairest, but this was in a gateway scenario. C2 also found that FPP was the least

preferred fairness definition. This can be interpreted to mean that participants understood the experiment well, as the unqualified would have been invited at the expense of the qualified. This interpretation is also consistent with the fairness calculus in selection decisions. It also suggests that participants considered the business utility of the selection procedure. The FPP was included as a fairness metrics because it can potentially reduce workplace discrimination by detecting phenomena such as in-group favoritism and nepotism (e.g. offering resources to family and friends), when this false-positive rate, the share of unqualified which were selected, is not equal across demographic groups.

Harms that arise from AI/ML through AI/ML having agency over decisions that affects humans (e.g. resume pre-screening) represent one category of AI/ML induced harms in responsible AI/ML. Another aspect of responsible AI/ML is the category of indirect, not-obvious harms that arise from human-machine interaction with AI/ML. For instance, when students rely heavily on LLMs for programming tasks, they might not develop essential problem-solving skills. This outsourcing of cognitive activities could lead to a dependency on AI tools and a decrease in critical thinking abilities. At the same time, AI tools also could help students and provide assistance in their learning. GenAI tutors also provide hints and learning suggestions. Therefore, the question is how to mitigate the effect of cognitive outsourcing and offering learning gains from GenAI.

For this aim C3 introduced a self-selected fixed-faded scaffolding intervention for the GenAI tutor, aiming to promote critical thinking through regulated feedback. This approach encouraged students to engage in deep thinking and self-explanation, fostering problem-solving and transfer skills (Chi et al., 1989; Schworm & Renkl, 2007). The study in C3 involved two groups that had either unrestricted access to a GenAI tutor and a group where the access was restricted by 90 seconds waiting timer after feedback was requested. Contrary to expectations, the group with full-access to feedback performed as well or better in transfer tasks compared to the control group, challenging the assumption that forced self-reflection improves problem-solving capabilities. A possible reason is insufficient mastery among learners despite individual efforts.

In C4 emergent behavior of complex ADM-ML systems is studied when user exhibiting agency over the ADM-ML system to degree that vary on the predicted outcome of the ADM-ML system and the characteristics of the user. The so-called right-to-be-forgotten (RBF) grants users control rights over their personal data and organizations must implement adequate data handling strategies that comply with privacy legislation. In the simulation study, C4 it was

found that certain strategies have undesirable side-effects. The strategy of masking the records belonging to the user increased unfairness, whereas data deletion did not lead to an increase in unfairness. The latter finding is remarkable, as it was hypothesized that deleting data from minority groups – as was the scenario in the simulation study – through RBF decrease the model's performance to accurately predict records belonging to the same group as the training data size is reduced and typically larger training data improves prediction quality. Remarkably, it was found that strategies that seek to maintain data availability through masking/imputation decreased the model performance unevenly for the younger group. Therefore, the trade-off between complying with the privacy requirement (and discarding all information about the record) and utility/fairness considerations must be balanced (e.g., utilizing a masking strategy that acknowledges, e.g., the age range). It also was found for most data handling strategies under investigation the more deletions were requested and addressed through one of the data handling strategies (not necessarily leading to record deletion) unfairness increased. Therefore, organization may be advised to take measures that mitigate the extend of data deletion requests which are accompanied by appropriate data handling strategies (data deletion).

Besides RBF organizations must consider privacy legislations also in general when handling and processing data. So-called privacy-preserving data sharing strategies allows organizations greater flexibility in handling and processing customer data including for purposes such as data analytics. Privacy-preserving data sharing strategies (PPDS) alter the original data to protect privacy. At the same time these strategies seek to maintain overall utility of the dataset as an organization usually is only interested in the overall utility of the dataset rather than individual records. In C5 it was found that a data science pipeline, which prepares customer data for sharing with internal/external data analysts, requires multiple iterations of testing PPDS algorithms and subsequent analysis to optimize PPDS algorithms not only for privacy but fairness as well, considering that PPDS algorithms can increase and decrease fairness. This finding implies that PPDS do no necessary decrease/increase fairness of ML models that are trained on privacy-preserving datasets rather than the original datasets. However, C5 found that overall performance is reduced with PPDS but this loss in performance does not necessarily lead to a loss in fairness.

C4-C5 explore how unfairness in complex ML systems may or may not arise. Whereas C1 and C2 explore how to link a technical metric to a psychological perceived factor when humans describe a ML system as unfair. What these contributions have in common is that ML is used to make decisions. These decisions lead to unfairness when the decisions are interlinked with benefits. The unfairness materializes through the allocation of benefits through ML-ADM when

the probability of (not) receiving the benefit varies significantly over belongingness to a demographic group conditional on eligibility. In a fair situation that conditional probability should be independent on group belonging. In contrast, C3 explores other harms that can arise from using ML/AI-based systems. In particular, it studies harms that arise from hybrid human-AI work when using LLM to support cognitive tasks. The types of harms which can occur in using ML/AI are summarized in Table 10.1.

Table 10.1. Harm-types in responsible ML/AI

AI/ML harms	**Contributions**
Decision-induced harms	C1-C2, C4-C5
Other harms	C3

Overall, C1-C3 address RO1a by investigating the relationship between humans and AI/ML systems empirically. These contributions investigate how AI/ML systems impact humans or are perceived by humans. C4 and C5 addresses RO1b by studying ML in complex systems that consists of multiple actors and where emergent behavior arises when multiple actors are present. C3 & C4 study technical sources of unfairness whereas C1 & C2 the link to human perception of technical unfairness. Together, C1-C5 make contributions to the knowledge base on responsible AI/ML in ISR.

In C6, the RO2 was addressed regarding the question of how organizations can derive value from ML systems by developing a system for supporting supplier-sourcing in industrial procurement. For this, a prescriptive analytics solution was developed that considers the specifics of the procurement process (e.g., the requirement to obtain at least one supplier offer and biased training data). For this, several techniques were explored. Interestingly, it was found that a dynamic rule, one that uses the information from every obtained supplier offer to update the price distribution, prescribes further searches for suppliers when there are large, expected cost savings in the direct procurement costs. In addition, the dynamic technique also handles biases in the training data better and can adjust itself. It is also found that even though the underlying parameters of the distribution are estimated with a noticeable error (high MAPE), this error does not translate in large decision errors as the developed system is able to balance procurement process and direct costs. In addition, the approximation using the Gamma distribution of the empirical data appears to be robust. The dynamic rule that uses Bayesian updating also shows robustness against deficient forecasts, which are deficient estimates are corrected by exploring more supplier offers.

C7 extends C6 to a more general framework using a different learning strategy. While C6 modelled the costs explicitly, C7 models the costs implicitly by using a deep neural network that is trained from its decision-making and using a gradient-free MOO approach based on PSO. The contribution offers thus an alternative to the *predict-first-then-optimize paradigm* by optimizing and predicting simultaneously.

In C8 the approach from C7 was applied to another learning paradigm and use-case. A specific feature engineering for handling financial time series was used. Based on these features, movements of the underlying financial time series were predicted. Instead of a PSO an evolutionary optimization approach was used for the single-objective case of maximizing trading profitability of the obtained trading rules. In the contribution also a specific mutation strategy was tested. A low-controlled adaptive mutation strength mechanism appears to perform better than a more aggressive mutation strategy in the studied setting of optimizing thresholds. Although in general local optima are a relevant problem for the class of heuristic optimization strategies, not reaching the local optima appears to be the reason for the subpar performance of the proposed aggressive mutation strategy. This mutation strategy replaces offsprings that are too close to their parents with randomly initialized values, yet this approach is subpar for the studied setting as the (local) optima are located on narrow peaks that are difficult to reach with the aggressive mutation strategy but could be reached with a gradual strategy (low-controlled).

In C9, a problem specific encoding strategy was used accounting for the properties of a spatial bike-sharing network. Also, spatial cross-validation procedures, which are novel to the IS literature, were introduced to model the problem of adding or removing stations from a bike-sharing station network. In particular, the use of spatial cross-validation can improve the predictive performance of the ML system because the ML models optimal during tuning tend to have better external validity. The newly proposed feature variables (vicinity/target) used to model the spatial network station facilitate better planning decisions. An interactive planning map was implemented as a design artifact to assist planners in an intuitive and efficient manner. The application can also be used, e.g., in combination with crowdsourcing approaches (Zhang et al., 2016; He & Kang G. Shin, 2018), where the interactive planning map application may facilitate acceptance and transparency of planning decisions.

Overall, C6-C9 make distinct contributions on the knowledgebase of ML in ISR.

10.2 Theoretical Implications

The research in this cumulative thesis has contributed to theory in multiple ways. Contributions C3, C4, and C5 provide descriptive and normative knowledge about designing *responsible*

AI/ML systems. Contributions C4 and C5 directly study how to balance privacy and fairness. While C4 focuses mainly on the record level, C5 focuses on the dataset level. Contribution C4 enhances understanding of dynamics in complex ML systems and the emergence of feedback loops that cause unfairness. Contribution C5 benchmarks privacy protection methods for data dissemination. Contribution 4 provides recommendations regarding their properties for maintaining privacy, utility, and fairness. Overall, these contributions improve the understanding of unintended consequences of AI/ML in complex organizational systems, contributing to the knowledge base of the dark sides of IS in ISR. Contribution C3 studies the potential consequences of AI on knowledge workers' cognitive processes and proposes an interface design to mitigate negative consequences such as skill-erosion. These contributions enhance understanding in ISR of what comprises responsible governance of AI/ML systems (Papagiannidis et al., 2025). Contributions C6-C9 provide descriptive knowledge regarding the procedural design of *value-oriented AI/ML* in organizations. These contributions deliver conceptual knowledge about how organizations can build AI/ML systems that align with their business objectives. Taking various technical forms, they are based on the underlying principle of value generation in the organizational use of AI/ML. These contributions describe implementation details and thus add *procedural knowledge* to the AI/ML knowledge base in ISR. The technical contributions comprise *encodings* (C9), *learning strategies* (C7), and *algorithmic systems* (C6, C8). Other researchers can build on this procedural knowledge by either extending it or applying it to other application domains, expanding the AI/ML knowledge base in ISR. Contributions C6 and C9 also identify application domains of AI/ML in organizational decision-making. These contributions describe how AI/ML can be used to influence business outcomes and how the application environment affects system design. They further solidify the understanding in ISR of AI/ML's role in organizations. Contribution C6 identifies a business process (sourcing of supplier quotes) that is supported by AI/ML and conceptualizes it into a quantitative form solvable with AI/ML applications. This conceptualization provides a basis for other academic contributions to building AI/ML systems in procurement. Contribution C9 studies the use of AI/ML in redesigning existing bike-sharing networks.

Table 10.2. Contributions to Value-Oriented AI/ML in Organizations using the terminology of Samtani et al. (2023)

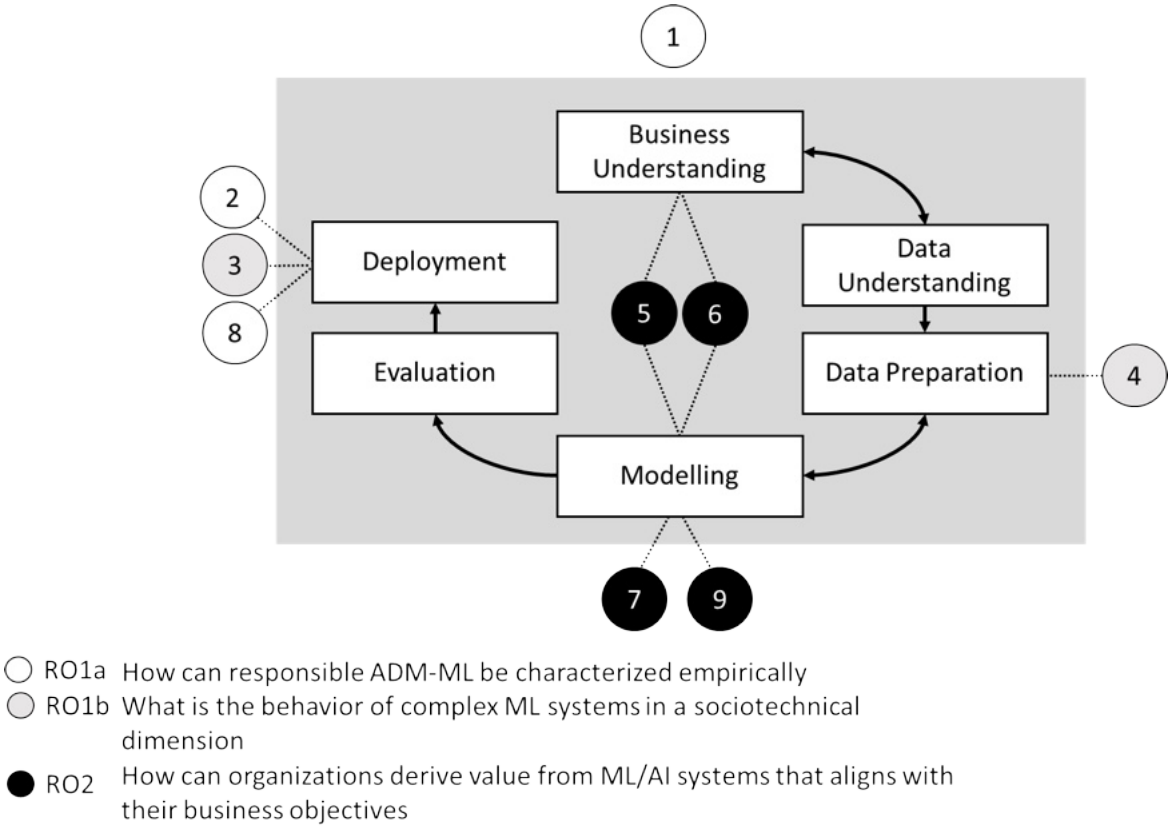
	Type	Domain	Function	Description
C6	Framework	Industry	Sourcing	• Framework for sourcing optimization of electronic resistors • Modeling approach via Bayesian optimal stopping • Exploration vs. exploitation trade-off
C7	Learning	General	Operations Management	• Multi-objective Optimization of DNN weights using PSO algorithm • Simulates business consequences rather than using conventional loss functions
C8	Domain-Specific	FinTech	Trading	• Combine rule-mining and heuristic optimization • Interpretable representation • Trading strategy with multi-horizon planning
C9	Repres-entation	Bike-Sharing	Location Planning	• Encoding spatial network (target areas) using weight matrices • Comparing spatial ML with conventional ML

10.3 Practical Implications

This thesis also provides several practical implications. RO1 discusses how to design AI/ML systems such that they are responsible. RO1 includes addressing issues such as bias, discrimination, and privacy concerns, which are critical for ensuring that AI/ML systems are socially sustainable, uphold an organization's reputation, and comply with regulations (Enholm et al., 2022). Compliance with regulations is also forward-looking, as human-facing AI/ML systems undergo social evaluation after deployment, which shapes societal norms and regulations regarding those systems (Papagiannidis et al., 2025). Therefore, the responsible use

of AI/ML systems is inherently in organizations' interests. AI governance, to which RO1 contributes, functions as the organizational mechanism ensuring that deployed AI/ML systems operate sustainably (Enholm et al., 2022; Papagiannidis et al., 2025). RO1 thus has addressed through empirical and algorithmic research the conceptualization of AI governance for the practice of AI/ML in organizations.

Figure 10.2. Contributions are embedded along the CRISP-DM framework according to their primary contribution.
(adapted from Wirth et al., 2000 and inspired by a similar figure in Stahmann 2024 p.23)



RO2 made practical contributions towards value-oriented AI/ML (Enholm et al., 2022). In particular, while considerable technological advances in AI/ML have been made in computer science, ISR makes practical contributions at the interface of technology and how technology affects societal and business outcomes (Samtani et al., 2023). The value-oriented implications as well as the implications regarding AI governance are discussed further using the CRISP-DM (Wirth et al., 2000). The CRISP-DM is a practical framework entailing all phases of building ML applications in practice. Therefore, I consider this framework suitable for presenting the practical contributions this thesis has made. While all practical contributions have used all phases of the CRISP-DM, the main implications of these contributions are discussed in specific phases of the CRISP-DM. Contributions 6 and 7 are rooted in business understanding and

translating this business understanding into an algorithmic solution. Contribution 4 benchmarks data pre-processing methods and thus provides practical recommendations for the data preparation phase in the CRISP-DM. Contributions 6-7 and 9 provide concrete modeling recommendations that can be used in practice. These contributions range from application-specific design implementations (Contribution 6, Contribution 9) to more general applicable learning strategies for operations management tasks (Contribution 7). Finally, Contributions 2-4 make recommendations regarding the deployment of AI/ML. For instance, Contribution 2 makes recommendations regarding designing AI/ML solutions for resume pre-screening in a job application context where selections are considered fair. Contribution 4 raises awareness for the problem of fairness-aware implementations of the RBF, which has important implications for data management practice.

1.4 Limitations

As with any research, this thesis also needs to acknowledge some limitations. First, I have only investigated a limited set of issues arising from RO1 and RO2. For instance, other sources of unfairness other than privacy are not considered. Also, responsible AI/ML is conceptualized in the aspects of privacy, fairness, and skill-erosion. Other topics, such as transparency or broader sociological phenomena (technology-induced loneliness), were not studied. The scope of the thesis is thus limited. Second, how organizations can derive value from AI/ML was only discussed in some specific application domains. However, the specificity of some application domains, such as procurement sourcing, inspired specific algorithmic solutions that can be of value for other domains.

11. Conclusion

In this cumulative thesis, I have investigated the challenges that organizations face when implementing AI/ML for their business processes. In nine contributions, this thesis provides an understanding of the sociotechnical system of AI/ML in organizations, from the design of responsible AI/ML systems to the design of value-oriented AI/ML systems. The thesis extends the knowledge base in ISR about the design of responsible AI/ML systems by conducting empirical research on the human-reaction towards exposure to AI/ML systems. Furthermore, it conducts algorithmic research (simulation studies) about emergent properties of complex ML systems, including effects on the human counterparts. In particular, technical sources of ML unfairness were investigated. Overall, the thesis derived knowledge about how to design such systems to form a synergetic relationship between human and AI/ML. Organizations implement responsible AI practices not solely out of altruism, but as a result of strategic considerations to mitigate public backlash and to proactively comply with evolving regulatory frameworks. The findings and research approaches presented in this thesis serve not only private entities but also regulatory bodies seeking to exercise appropriate oversight of AI systems and address potential adverse impacts before they materialize in organizational settings. Congruent with the instrumental objective of responsible AI, the thesis also provides insights into how organizations can derive value from AI/ML. Four further contributions present artifacts developed to enhance organizational decision-making and novel algorithmic solutions to operational problems in organizations. Maintaining a synergetic relationship between human and AI/ML is relevant for organization to maintain a benevolent relationship. The contributions made in this thesis are therefore relevant for both theory and practice, as they provide insights into the challenges and opportunities of AI/ML in organizations.

Index

References

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2). https://doi.org/10.17705/1jais.00423

Afzalan, N., & Sanchez, T. (2017). Testing the use of crowdsourced information: Case study of bike-share infrastructure planning in cincinnati, ohio. In *Urban Planning* (No. 3; Vol. 2, pp. 33–44). https://doi.org/10.17645/up.v2i3.1013

Agency for Geoinformation and Surveying. (2014). *ALKIS verwaltungsgrenzen hamburg*. Published by *Ministry of Urban Development and Housing of the Free and Hanseatic City of Hamburg*.

Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J. S. Z., Jarvenpaa, S., Ram, S., & Mikalef, P. (2022). Artificial Intelligence in Information Systems: State of the Art and Research Roadmap. *Communications of the Association for Information Systems*, *50*(1), pp. 420–438. https://doi.org/10.17705/1CAIS.05017

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning*. Published by *PMLR*.

Albach, M., & Wright, J. R. (2021). The role of accuracy in algorithmic process fairness across multiple domains. *Proceedings of the 22nd ACM Conference on Economics and Computation*. https://doi.org/https://doi.org/10.1145/3465456.3467620

Albright, S. C. (1977). A bayesian approach to a generalized house selling problem. *Management Science*, *24*(4), pp. 432–440. https://doi.org/10.1287/mnsc.24.4.432

Alcorn, L. G., & Jiao, J. (2019). Bike-sharing station usage and the surrounding built environments in major texas cities. *Journal of Planning Education and Research*, *84*(3), pp. 0739456X1986285. https://doi.org/10.1177/0739456X19862854

Allal-Chérif, O., Simón-Moya, V., & Ballester, A. C. C. (2021). Intelligent purchasing: How artificial intelligence can redefine the purchasing function. *Journal of Business Research*, *124*, pp. 69–76. https://doi.org/10.1016/j.jbusres.2020.11.050

Alter, S. (2008). Defining information systems as work systems: Implications for the IS field. *European Journal of Information Systems*, *17*(5), pp. 448–469. https://doi.org/10.1057/ejis.2008.37

Alves, W. M., & Rossi, P. H. (1978). Who Should Get What? Fairness Judgments of the Distribution of Earnings. *American Journal of Sociology*, *84*(3), pp. 541–564. https://doi.org/10.1086/226826

Ambrose, M. L., & Schminke, M. (2009). The role of overall justice judgments in organizational justice research: A test of mediation. *Journal of Applied Psychology*, *94*(2), pp. 491–500. https://doi.org/10.1037/a0013203

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine bias. *ProPublica*.

Anselin, L., & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers in Regional Science*, *65*(1), pp. 11–34.

Araujo, T., Helberger, N., Kruikemeier, S., & Vreese, C. H. de. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), pp. 611–623.

Arnott, D., & Pervan, G. (2005). A Critical Analysis of Decision Support Systems Research. *Journal of Information Technology*, *20*(2), pp. 67–87. https://doi.org/10.1057/palgrave.jit.2000035

Arnott, D., & Pervan, G. (2016). A Critical Analysis of Decision Support Systems Research Revisited: The Rise of Design Science. In L. P. Willcocks, C. Sauer, & M. C. Lacity (Eds.), *Enacting Research Methods in Information Systems: Volume 3* (pp. 43–103). Published by *Springer International Publishing*. https://doi.org/10.1007/978-3-319-29272-4_3

Arnott, D., Pervan, G., & Curtin University. (2012). Design Science in Decision Support Systems Research: An Assessment using the Hevner, March, Park, and Ram Guidelines. *Journal of the Association for Information Systems*, *13*(11), pp. 923–949. https://doi.org/10.17705/1jais.00315

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

Austrian data protection authority (2018). *DSB Bescheid 5.12.2018, D123.270/0009-DSB/2018 – Löschen durch Anonymisieren*. https://www.ris.bka.gv.at/Dokumente/Dsk/DSBT_20181205_DSB_D123_270_0009_DSB_2018_00/DSBT_20181205_DSB_D123_270_0009_DSB_2018_00.pdf

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), pp. 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019a). Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019b). Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

Bao, J., Shi, X., & Zhang, H. (2018). Spatial analysis of bikeshare ridership with smart card and POI data using geographically weighted regression method. *IEEE Access*, *6*, pp. 76049–76059. https://doi.org/10.1109/ACCESS.2018.2883462

Barbieri, C. A., Booth, J. L., Begolli, K. N., & McCann, N. (2021). The effect of worked examples on student learning and error anticipation in algebra. *Instructional Science*, *49*(4), pp. 419–439. https://doi.org/https://doi.org/10.1007/s11251-021-09545-6

Barocas, S., & Selbst, A. D. (2016). Big data has disparate impact. *California Law Review*, *104*(3), pp. 671–732. https://doi.org/10.2139/ssrn.2477899

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, *34*(4), pp. 1582–1602. https://doi.org/10.1287/isre.2023.1199

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias* (No. arXiv:1810.01943). Published by *arXiv*. https://doi.org/10.48550/arXiv.1810.01943

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis. *Review of Educational Research*, *87*(2), pp. 309–344. https://doi.org/10.3102/0034654316670999

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*, *13*(2), pp. 273–291. https://rjournal.github.io/

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), pp. 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Berkel, N. van, Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), pp. 1–21. https://doi.org/10.1145/3359130

Berkel, N. van, Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of information presentation on fairness perceptions of machine learning predictors. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI conference on human factors in computing systems*. Published by *Association for Computing Machinery*. https://doi.org/10.1145/3411764.3445365

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*. https://doi.org/10.1287/mnsc.2018.3253

Bieliński, T., Kwapisz A., & Ważna, A.. (2019). Bike-sharing systems in poland. *Sustainability*, *11*(9), pp. 1–14. https://doi.org/10.3390/su11092458

Bienhaus, F., & Haddud, A. (2018). Procurement 4.0: Factors influencing the digitisation of procurement and supply chains. *Business Process Management Journal*, *24*(4), pp. 965–984. https://doi.org/10.1108/BPMJ-06-2017-0139

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. https://doi.org/10.1145/3173574.3173951 pp.1–14

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (MSR-TR-2020-32).

Published by *Microsoft*. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
Bishop, L. (2023, January 26). *A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing* (SSRN Scholarly Paper No. 4338981).
Biwer, F., Egbrink, M. G. A. oude, Aalten, P., & de Bruin, A. B. H. (2020). Fostering Effective Learning Strategies in Higher Education – A Mixed-Methods Study. *Journal of Applied Research in Memory and Cognition*, *9*(2), pp. 186–203. https://doi.org/10.1016/j.jarmac.2020.03.004
Boer, L. de, Harink, J., & Heijboer, G. (2002). A conceptual model for assessing the impact of electronic procurement. *European Journal of Purchasing &Amp; Supply Management*, *8*(1), pp. 25–33. https://doi.org/10.1016/s0969-7012(01)00015-6
Boute, R. N., Gijsbrechts, J., Jaarsveld, W. van, & Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research*, *298*(2), pp. 401–412. https://doi.org/10.1016/j.ejor.2021.07.016
Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), pp. 324. https://doi.org/10.2307/2334029
Brand, R. (2002). *Microdata protection through noise addition*. pp. 97–116. https://doi.org/10.1007/3-540-47804-3_8
Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, *5*(6), pp. 853–862. https://doi.org/10.5194/nhess-5-853-2005
Brenning, A. (2012). *Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The r package sperrorest: 2012 IEEE international geoscience and remote sensing symposium*. https://doi.org/10.1109/IGARSS.2012.6352393
Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Published by *The Belknap Press of Harvard University Press*. https://doi.org/10.4159/9780674419377
Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), pp. 1–21. https://doi.org/10.1145/3449287
Bürgeramt, Statistik und Wahlen der Stadt Frankfurt am Main. (2017). *Wahlgebietsgliederung 2017*. http://offenedaten.frankfurt.de/dataset/a08fa6dc-8d5e-4881-a0cb-b59504165a60/resource/099ad3e8-96eb-4b5f-91a2-a783587509e7/download/wahlbezirke2017.shp.zip
Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), pp. 290–298. https://doi.org/10.1037/a0031026
Cai, Z., Mao, P., Wang, D., He, J., Chen, X., & Fan, X. (2022). Effects of Scaffolding in Digital Game-Based Learning on Student's Achievement: A Three-Level Meta-analysis. *Educational Psychology Review*, *34*(2), pp. 537–574. https://doi.org/10.1007/s10648-021-09655-0
Cao, G., Jin, G. Z., Weng, X., & Zhou, L.-A. (2018). Market expanding or market stealing? Competition with network effects in BikeSharing. In *Working Paper Series*. Published by *National Bureau of Economic Research*. https://doi.org/10.3386/w24938
Carbonneau, G. A. V., R And Kersten. (2008). Predicting opponent's moves in electronic negotiations using neural networks. *Expert Systems with Applications*, *34*(2), pp. 1266–1273. https://doi.org/10.1016/j.eswa.2006.12.027
Carbonneau, R. A., Kersten, G. E., & Vahidov, R. M. (2011). Pairwise issue modeling for negotiation counteroffer prediction using neural networks. *Decision Support Systems*, *50*(2), pp. 449–459. https://doi.org/10.1016/j.dss.2010.11.002
Carvalho, T., Moniz, N., & Antunes, L. (2023a). A three-way knot: privacy, fairness, and predictive performance dynamics. In *EPIA Conference on Artificial Intelligence* (pp. 55-66). Springer. https://doi.org/10.1007/978-3-031-49008-8_5
Carvalho, T., Moniz, N., Faria, P., & Antunes, L. (2023b). Survey on privacy-preserving techniques for microdata publication. *ACM Computing Surveys*, 55(14s), 1-42. https://doi.org/10.1145/3588765
Caton, S., Malisetty, S., & Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research*, *74*, pp. 1011–1035. https://doi.org/10.1613/jair.1.13197

Chai, J., & Ngai, E. W. T. (2020). Decision-making techniques in supplier selection: Recent accomplishments and what lies ahead. *Expert Systems with Applications*, *140*, pp. 112903. https://doi.org/10.1016/j.eswa.2019.112903

Chamikara, M., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2020). Efficient privacy preservation of big data for accurate data mining. *Information Sciences*, *527*, pp. 420–443. https://doi.org/10.1016/j.ins.2019.05.053

Chang Chien, Y.-W., & Chen, Y.-L. (2010). Mining associative classification rules with stock trading data – a GA-based method. *Knowledge-Based Systems*, *23*(6), pp. 605–614. https://doi.org/https://doi.org/10.1016/j.knosys.2010.04.007

Chang, H., & Shokri, R. (2021). *On the privacy risks of algorithmic fairness*.

Chen, H., Zhu, T., Zhang, T., Zhou, W., & Yu, P. S. (2023). Privacy and fairness in federated learning: On the perspective of tradeoff. *ACM Computing Surveys*, *56*, pp. 1–37. https://doi.org/10.1145/3606017

Chen, J. K. T., Valliant, R. L., & Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), pp. 657–681. https://doi.org/10.1111/rssc.12327

Chen, L., Zhang, D., Pan, G., Ma, X., Yang, D., Kushlev, K., Zhang, W., & Li, S. (2015). Bike sharing station placement leveraging heterogeneous urban open data. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 571–575. https://doi.org/10.1145/2750858.2804291 pp.571–575

Chen, Q., & Sun, T. (2015). A model for the layout of bike stations in public bike-sharing systems. *Journal of Advanced Transportation*, *49*(8), pp. 884–900. https://doi.org/https://doi.org/10.1002/atr.1311

Chen, Z., & Rossi, R. (2021). A dynamic ordering policy for a stochastic inventory problem with cash constraints. *Omega*, *102*, pp. 102378. https://doi.org/10.1016/j.omega.2020.102378

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., & Zhang, Y. (2021, November). When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security* (pp. 896-911). https://doi.org/10.1145/3460120.3484756

Cheng, H.-F., Stapleton, L., Wang, R., Bullock, P., Chouldechova, A., Wu, Z. S., & Zhu, H. (2021). Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI conference on human factors in computing systems*. Published by *Association for Computing Machinery*. https://doi.org/10.1145/3411764.3445308

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), pp. 145–182. https://doi.org/10.1207/s15516709cog1302_1

Choudhury, V., Hartzel, K. S., & Konsynski, B. R. (1998). Uses and consequences of electronic markets: An empirical investigation in the aircraft parts industry. *MIS Quarterly*, *22*(4), pp. 471. https://doi.org/10.2307/249552

Christensen, R. H. B. (2018). *Cumulative Link Models for Ordinal Regression with the R Package ordinal*. https://cran.r-universe.dev/ordinal/doc/clm_article.pdf

Cintrano, C., Chicano, F., Stützle, T., & Alba, E. (2018). Studying solutions of the p-median problem for the location of public bike stations. In F. Herrera, S. Damas, R. Montes, S. Alonso, Ó. Cordón, A. González, & A. Troncoso (Eds.), *Advances in artificial intelligence* (pp. 198–208). Published by *Springer International Publishing*. https://doi.org/10.1007/978-3-030-00374-6_19 pp.198–208

Conde S., D. J., Kämpf, N. L., Rößler-von Saß, D., Schurig, T. & Kliewer, N., "Privacy-Preserving Data Sharing: A Systematic Review and Future Research Areas" (2024). *ECIS 2024 Proceedings*. 12. https://aisel.aisnet.org/ecis2024/track10_dmds_ecosystems/track10_dmds_ecosystems/12

Coello, C. C., & Lechuga, M. S. (2002). MOPSO: A proposal for multiple objective particle swarm optimization. *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02*, pp. 1051–1056. https://doi.org/10.1109/cec.2002.1004388 pp.1051–1056

Colquitt, J. A., & Shaw, J. C. (2005). How should organizational justice be measured? In *Handbook of organizational justice* (pp. 113–152). Published by *Lawrence Erlbaum Associates Publishers*. https://psycnet.apa.org/record/2005-03594-004

Cowgill, B., Dell'Acqua, F., & Matz, S. (2020). The managerial effects of algorithmic fairness activism. *AEA Papers and Proceedings*, *110*, pp. 85–90. https://doi.org/10.1257/pandp.20201035

Crama, Y., Pascual J, R., & Torres, A. (2004). Optimal procurement decisions in the presence of total quantity discounts and alternative product recipes. *European Journal of Operational Research*, *159*(2), pp. 364–378. https://doi.org/10.1016/j.ejor.2003.08.021
Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Review*, *55*, pp. 93–128. https://doi.org/10.2139/ssrn.2477899
Croci, E., & Rossi, D. (2014). Optimizing the position of bike sharing stations. The milan case. In *SSRN Electronic Journal*. Published by *CERE Working Paper, 68*. https://doi.org/10.2139/ssrn.2461179
Cui, R., Li, M., & Zhang, S. (2021). AI and procurement. *Manufacturing & Service Operations Management*, *24*(2), pp. 691–706. https://doi.org/10.1287/msom.2021.0989
Cui, R., Li, M., & Zhang, S. (2022). AI and procurement. *Manufacturing &Amp; Service Operations Management*, *24*(2), pp. 691–706. https://doi.org/10.1287/msom.2021.0989
Cummings, R., Gupta, V., Kimpara, D., & Morgenstern, J. (2019). *On the compatibility of privacy and fairness*. pp. 309–315. https://doi.org/10.1145/3314183.3323847
Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023, April). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. https://doi.org/10.35542/osf.io/hcgzj
Dastin, J. (2018). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. In *Reuters*. Published by *https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/*.
De Moor, B. J., Gijsbrechts, J., & Boute, R. N. (2022). Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. *European Journal of Operational Research*, *301*(2), pp. 535–545. https://doi.org/10.1016/j.ejor.2021.10.045
Dennis, A. R., & Valacich, J. S. (2001). Conducting Experimental Research in Information Systems. *Communications of the Association for Information Systems*, *7*. https://doi.org/10.17705/1cais.00705
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), pp. 1155–1170. https://doi.org/10.1287/mnsc.2016.2643
Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Inf. Syst. Res.*, *17*, pp. 61–80. https://doi.org/10.1287/isre.1060.0080
Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 275–285). Published by *Association for Computing Machinery; ACM*. https://doi.org/10.1145/3301275.3302310 pp.275–285
Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), pp. 754–818. https://doi.org/10.1111/isj.12370
Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. Lane, J. Theeuwes, & L. Zayatz (Eds.), *Confidentiality, disclosure and data access: Theory and practical applications for statistical agencies* (pp. 111–134). Published by *Elsevier*.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. https://doi.org/10.1145/2090236.2090255 pp.214–226
Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). *Calibrating noise to sensitivity in private data analysis*. pp. 265–284. https://doi.org/10.1007/11681878_14
Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing* (2nd ed.). Published by *Springer*. https://doi.org/https://doi.org/10.1007/978-3-662-44874-8
El-Assi, W., Salah Mahmoud, M., & Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: A station level analysis of commercial bike sharing in toronto. *Transportation*, *44*(3), pp. 589–613. https://doi.org/10.1007/s11116-015-9669-z
Electronics, M. of. (2018). *& information technology, government of india*. https://www.meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf
Eliot, D. L. B. (2021). Deskilling Of Lawyers Due To AI Is Not A Foregone Conclusion. In *SSRN Electronic Journal* ({{SSRN Scholarly Paper}} No. 3956781). Published by *Elsevier BV*. https://doi.org/10.2139/ssrn.3956781

Elliot, M., Hundepool, A., Nordholt, E. S., Tambay, J.-L., & Wende, T. (2005). *Glossary on statistical disclosure control*. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.45.e.pdf?msclkid=098acdbfb8d811ecab1f4255d0b86f8b

Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize." *Management Science*, *68*. https://doi.org/10.1287/mnsc.2020.3922

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*, *24*(5), pp. 1709–1734. https://doi.org/10.1007/s10796-021-10186-w

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (Vol. 81, pp. 160–171). Published by *PMLR*. https://proceedings.mlr.press/v81/ensign18a.html pp.160–171

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in montreal. *Journal of Transport Geography*, *41*, pp. 306–314. https://doi.org/10.1016/j.jtrangeo.2014.01.013

Farshid, S., Reitz, A., & Roßbach, P. (2019). Design of a forgetting blockchain: A possible way to accomplish GDPR compatibility. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 1–9. https://doi.org/10.24251/HICSS.2019.850 pp.1–9

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. **Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia: Association for Computing Machinery*, pp. 259–268. https://doi.org/10.1145/2783258.2783311 pp.259–268

Ferreira, K. J., Simchi-Levi, D., & Wang, H. (2018). Online network revenue management using thompson sampling. *Operations Research*, *66*(6), pp. 1586–1602. https://doi.org/10.1287/opre.2018.1755

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and Opportunities. *Business & Information Systems Engineering*, *62*(4), pp. 379–384. https://doi.org/10.1007/s12599-020-00650-3

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, *36*(1), pp. 92–113. https://doi.org/10.1080/01441647.2015.1033036

Frade, I., & Ribeiro, A. (2015). Bike-sharing stations: A maximal covering location approach. *Transportation Research Part A: Policy and Practice*, *82*, pp. 216–227. https://doi.org/10.1016/j.tra.2015.09.014

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *10/12/2015*, pp. 1322–1333. https://doi.org/10.1145/2810103.2813677

Fricker, C., & Gast, N. (2016). Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity. *EURO Journal on Transportation and Logistics*, *5*(3), pp. 261–291. https://doi.org/10.1007/s13676-014-0053-5

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338. https://doi.org/10.1145/3287560.3287589 pp.329–338

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing. *ACM Computing Surveys*, *42*(4), pp. 1–53. https://doi.org/10.1145/1749603.1749605

Gabel, T., & Riedmiller, M. (2011). Distributed policy search reinforcement learning for job-shop scheduling tasks. *International Journal of Production Research*, *50*(1), pp. 41–61. https://doi.org/10.1080/00207543.2011.571443

Gadde, L.-E., & Snehota, I. (2000). Making the most of supplier relationships. *Industrial Marketing Management*, *29*(4), pp. 305–316. https://doi.org/10.1016/s0019-8501(00)00109-7

Gajos, K. Z., & Mamykina, L. (2022). Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. *27th International Conference on Intelligent User Interfaces*, pp. 794–806. https://doi.org/10.1145/3490099.3511138 pp.794–806

Ganti, R. K., Pham, N., Tsai, Y.-E., & Abdelzaher, T. F. (2008). *PoolView*. pp. 281–294. https://doi.org/10.1145/1460412.1460440

García-Palomares, J. C., Gutiérrez, J., & Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, *35*(1), pp. 235–246. https://doi.org/10.1016/j.apgeog.2012.07.002

Gedye, S. (2010). Formative assessment and feedback: A review. *Planet*, *23*(1), pp. 40–45. https://doi.org/10.11120/plan.2010.00230040

Gehrke, S. R., & Welch, T. F. (2019). A bikeshare station area typology to forecast the station-level ridership of system expansion. *Journal of Transport and Land Use*, *12*. https://doi.org/10.5198/jtlu.2019.1395 Article No.1

Gel, E. S., & Salman, F. S. (2022). Dynamic ordering decisions with approximate learning of supply yield uncertainty. *International Journal of Production Economics*, *243*, pp. 108252. https://doi.org/10.1016/j.ijpe.2021.108252

Georg, A., Ivan, S., Florian, L., Claudia, W., & Markus, S. (2021). The FairCeptron: A framework for measuring human perceptions of algorithmic fairness. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*, pp. 135–140. https://doi.org/10.1145/3450614.3465224 pp.135–140

Ghasemaghaei, M., & Kordzadeh, N. (2025). Ethics in the Age of Algorithms: Unravelling the Impact of Algorithmic Unfairness on Data Analytics Recommendation Acceptance. *Information Systems Journal*, *n/a*(n/a). https://doi.org/10.1111/isj.12572

Ginart, A. A., Guan, M. Y., Valiant, G., & Zou, J. (2019). Making AI forget you: Data deletion in machine learning. In *Proceedings of the 33rd international conference on neural information processing systems*. Published by *Curran Associates Inc.* https://dl.acm.org/doi/10.5555/3454287.3454603

Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, *12*(2-3), pp. 304–325. https://doi.org/10.1080/17421772.2017.1300679

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pp. 903–912. https://doi.org/10.1145/3178876.3186138 pp.903–912

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2012, December 8). The case for process fairness in learning: Feature selection for fair decision making. *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*.

Griffin, G. P., & Jiao, J. (2019). Crowdsourcing bike share station locations. *Journal of the American Planning Association*, *85*(1), pp. 35–48. https://doi.org/10.1080/01944363.2018.1476174

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.* https://proceedings.mlr.press/v9/glorot10a.html

Guerreiro, T., Providelo, J., Pitombo, C., Ramos, R., & Rodrigues da Silva, Antônio Nélson. (2017). Data-mining, GIS and multicriteria analysis in a comprehensive method for bicycle network planning and design. *International Journal of Sustainable Transportation*. https://doi.org/10.1080/15568318.2017.1342156

Guida, M., Caniato, F., Moretto, A., & Ronchi, S. (2023). The role of artificial intelligence in the procurement process: State of the art and research agenda. *Journal of Purchasing and Supply Management*, *29*(2), pp. 100823. https://doi.org/10.1016/j.pursup.2023.100823

Gunarathne, P., Rui, H., & Seidmann, A. (2019). Racial discrimination in social media customer service: Evidence from a popular microblogging platform. **Proceedings of the 52nd Hawaii International Conference on System Sciences*, t. Bui (Ed.), Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2019.815

Günther, S. A. (2021). The impact of social norms on students' online learning behavior: Insights from two randomized controlled trials. *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 12–21. https://doi.org/10.1145/3448139.3448141 pp.12–21

Guo, C., Goldstein, T., Hannun, A., & Van Der Maaten, L. (2020). Certified data removal from machine learning models. *Proceedings of the 37th International Conference on Machine Learning*. https://doi.org/https://dl.acm.org/doi/10.5555/3524938.3525297

Guo, X., Yuan, Z., & Tian, B. (2009). Supplier selection based on hierarchical potential support vector machine. *Expert Systems with Applications*, *36*(3), pp. 6978–6985. https://doi.org/10.1016/j.eswa.2008.08.074

Guo, Y., Zhou, J., Wu, Y., & Li, Z. (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in ningbo, china. *PLOS ONE*, *12*(9), pp. e0185100. https://doi.org/10.1371/journal.pone.0185100

Haas, C. (2019). The price of fairness. A framework to explore trade-offs in algorithmic fairness. *ICIS 2019 Proceedings*, pp. 19. pp.19

Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook SpringerLink*. Published by *Springer Nature Switzerland AG*. https://doi.org/https://doi.org/10.1007/978-3-030-80519-7

Hamburger Verkehrsverbund GmbH. (2013). *HVV-haltestellennamen mit koordinate (hamburg)*. Published by *the Free and Hanseatic City of Hamburg*.

Handfield, R., Jeong, S., & Choi, T. (2019). Emerging procurement technology: Data analytics and cognitive analytics. *International Journal of Physical Distribution & Logistics Management*, *49*. https://doi.org/10.1108/IJPDLM-11-2017-0348

Hannan, J., Chen, H.-Y. W., & Joseph, K. (2021). Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 555–565. https://doi.org/10.1145/3461702.3462568 pp.555–565

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3323–3331. https://dl.acm.org/doi/10.5555/3157382.3157469 pp.3323–3331

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (2020). An empirical study on the perceived fairness of realistic, imperfect machine learning models. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 392–402. https://doi.org/10.1145/3351095.3372831

Hartman, R. (1972). The effects of price and cost uncertainty on investment. *Journal of Economic Theory*, *5*(2), pp. 258–266. https://doi.org/10.1016/0022-0531(72)90105-6

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), pp. 81–112. https://doi.org/10.3102/003465430298487

He, S., & Shin, K.. (2018). (Re)configuring bike station network via crowdsourced information fusion and joint optimization. In *Proceedings of the eighteenth ACM international symposium on mobile ad hoc networking and computing* (pp. 1–10). Published by *ACM*. https://doi.org/10.1145/3209582.3209583

Helberger, N., Araujo, T., & Vreese, C. H. de. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, pp. 105456. https://doi.org/10.1016/j.clsr.2020.105456

Herington, J. (2020). Measuring fairness in an unfair world. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3375627.3375854

Hobert, S. (2019). *Say Hello to 'Coding Tutor'! Design and Evaluation of a Chatbot-based Learning System Supporting Students to Learn to Program*.

Hobert, S., & Wolff, R. M. von. (2019). Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *Wirtschaftsinformatik 2019 Proceedings*.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Published by *Springer*. https://doi.org/10.1007/978-0-387-92407-6

Holstein, K., Vaughan, J. W., III, H. D., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300830

Hong, H., & Shum, M. (2006). Using price distributions to estimate search costs. *The RAND Journal of Economics*, *37*(2), pp. 257–275. https://doi.org/10.1111/j.1756-2171.2006.tb00015.x

Hu, S.-R., & Liu, C. T. (2014). An optimal location model for a bicycle sharing program with truck dispatching consideration: 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014. *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pp. 1775–1780. https://doi.org/10.1109/ITSC.2014.6957950

Huang, X., Wu, C., Du, X., Wang, H., & Ye, M. (2024). A novel stock trading utilizing long short term memory prediction and evolutionary operating-weights strategy. *Expert Systems with Applications*, *246*, pp. 123146. https://doi.org/https://doi.org/10.1016/j.eswa.2024.123146

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 49–58. https://doi.org/10.1145/3287560.3287600 pp.49–58

Hyland, M., Hong, Z., Pinto, Helen Karla Ramalho de Farias, & Chen, Y. (2018). Hybrid cluster-regression approach to model bikeshare station usage. *Transportation Research Part A: Policy and Practice*, *115*, pp. 71–89. https://doi.org/10.1016/j.tra.2017.11.009

Jahanshahi, D., Minaei, M., Kharazmi, O. A., & Minaei, F. (2019). Evaluation and relocating bicycle sharing stations in mashhad city using multi-criteria analysis. *International Journal of Transportation Engineering*, *6*(3), pp. 265–283.

Javier Garcia-Gutierrez, Javier Romero-Torres, & Juan Gaytan-Iniestra. (2014). Dimensioning of a bike sharing system (BSS): A study case in nezahualcoyotl, mexico. *Procedia - Social and Behavioral Sciences*, *162*, pp. 253–262. https://doi.org/10.1016/j.sbspro.2014.12.206

Jiawei Zhang, Xiao Pan, Moyin Li, & Philip S. Yu. (2016). Bicycle-sharing systems expansion: Station re-deployment through crowd planning. In *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 1–10). Published by *ACM*. https://doi.org/10.1145/2996913.2996926

Johnson, G. A., Shriver, S. K., & Du, S. (2020). Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, *39*(1), pp. 33–51. https://doi.org/10.1287/mksc.2019.1198

Jordon, J., Yoon, J., & Van Der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations*.

Kabak, M., Erbaş, M., Çetinkaya, C., & Özceylan, E. (2018). A GIS-based MCDM approach for the evaluation of bike-share stations. *Journal of Cleaner Production*, *201*, pp. 49–60. https://doi.org/10.1016/j.jclepro.2018.08.033

Kahraman, C., Cebeci, U., & Ulukan, Z. (2003). Multi-criteria supplier selection using fuzzy AHP. *Logistics Information Management*, *16*(6), pp. 382–394. https://doi.org/10.1108/09576050310503367

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*(3), pp. 209–215.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), pp. 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware Learning through Regularization Approach. *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. https://doi.org/10.1109/ICDMW.2011.83 pp.643–650

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*(3), pp. 224–232. https://doi.org/10.1198/000313006X124640

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, pp. 102274.

Kazemi, E., Zadimoghaddam, M., & Karbasi, A. (2018). Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. *International Conference on Machine Learning*, pp. 2544–2553. pp.2544–2553

Kellner, F., Lienland, B., & Utz, S. (2019). An a posteriori decision support methodology for solving the multi-criteria supplier selection problem. *European Journal of Operational Research*, *272*(2), pp. 505–522. https://doi.org/10.1016/j.ejor.2018.06.044

Khosrawi-Rad, B., Rinn, H., Schlimbach, R., Gebbing, P., Yang, X., Lattemann, C., Markgraf, D., & Robra-Bissantz, S. (2022, June 18). Conversational Agents in Education – A Systematic Literature Review. *ECIS 2022 Research Papers*. ECIS 2022, Munich, Germany. https://aisel.aisnet.org/ecis2022_rp/18 Article No.182

Kilincci, O., & Onal, S. A. (2011). Fuzzy AHP approach for supplier selection in a washing machine company. *Expert Systems with Applications*, *38*(8), pp. 9656–9664. https://doi.org/10.1016/j.eswa.2011.01.159

Kim, Y., Ngai, E. C.-H., & Srivastava, M. B. (2011). *Cooperative state estimation for preserving privacy of user behaviors in smart grid*. pp. 178–183. https://doi.org/10.1109/SmartGridComm.2011.6102313

Kizilcec, R. F. (2016). How much information? In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *CHI '16: CHI conference on human factors in computing systems*. Published by *The Association for Computing Machinery*.

Klepsch, M., & Seufert, T. (2021). Making an Effort Versus Experiencing Load. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.645284

Kloimüllner, C., & Raidl, G. R. (2017). Hierarchical clustering and multilevel refinement for the bike-sharing station planning problem. *International Conference on Learning and Intelligent Optimization*, pp. 150–165. pp.150–165. https://doi.org/10.1007/978-3-319-69404-7_11

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2021). Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context. *Business & Information Systems Engineering*, *63*(1), pp. 39–54. https://doi.org/10.1007/s12599-020-00673-w

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), pp. 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*, pp. 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*, pp. 5802–5805. https://doi.org/10.1073/pnas.1218772110

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), pp. 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *Eur. J. Oper. Res.*, *281*. https://doi.org/10.1016/j.ejor.2019.09.018

Krosnick, J. A., & Presser, S. (2010). Chapter 9: Question and questionnaire design. In P. Marsden & J. Wright (Eds.), *Handbook of survey research*. Published by *Emerald Group Publishing Limited*.

Kruse, R., Mostaghim, S., Borgelt, C., Braune, C., & Steinbrecher, M. (2022). *Computational swarm intelligence* (pp. 343–369). Published by *Springer International Publishing*.

Kuchar, J. (2015). *rCBA: CBA Classifier* (p. 0.4.3) [Data set]. Published by *Comprehensive R Archive Network*. https://doi.org/10.32614/CRAN.package.rCBA

Kuhn, M. (2015). Caret: Classification and regression training. *Astrophysics Source Code Library*, *1505*, pp. 003.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, pp. 116659. https://doi.org/10.1016/j.eswa.2022.116659

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in methods and practices in psychological science*, 1(2), 259-269. https://doi.org/10.1177/2515245918770963

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). Mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), pp. 1903. https://doi.org/10.21105/joss.01903

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), pp. 217–234. https://doi.org/10.1111/ijsa.12246

Larsen, K., & Bong, C.-H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly*, *40*, pp. 529–551; A1.

Lee, C. C., & Ou-Yang, C. (2009). A neural networks approach for forecasting the supplier's bid prices in supplier selection negotiation process. *Expert Systems with Applications*, *36*(2), pp. 2961–2970. https://doi.org/10.1016/j.eswa.2008.01.063

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), pp. 1–16. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, pp. 1035–1042. https://doi.org/10.1145/2998181.2998294

Lee, M. K., & Rich, K. (2021). Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445570

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, pp. 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003

LeSage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics* (Vol. 237). Published by *Chapman and Hall/CRC*. https://doi.org/10.1201/9781420064254

Li, N., Li, T., & Venkatasubramanian, S. (2007). *T-closeness: Privacy beyond k-anonymity and l-diversity*. pp. 106–115. https://doi.org/10.1109/ICDE.2007.367856

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, pp. 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lin, C.-T., Chen, C.-B., & Ting, Y.-C. (2011). An ERP model for supplier selection in electronics industry. *Expert Systems with Applications*, *38*(3), pp. 1760–1765. https://doi.org/10.1016/j.eswa.2010.07.102

Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, *97*, pp. 258–276. https://doi.org/10.1016/j.trc.2018.10.011

Lipnevich, A. A., & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.720195

Liu, Q., Deho, O., Vadiee, F., Khalil, M., Joksimovic, S., & Siemens, G. (2025). Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 591-600.

Liu, J., Li, Q., Qu, M., Chen, W., Yang, J., Xiong, H., Zhong, H., & Fu, Y. (2015). Station site optimization in bike sharing systems. *2015 IEEE International Conference on Data Mining*, pp. 883–888. https://doi.org/10.3390/ijgi10020062

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), pp. 410. https://doi.org/10.3390/educsci13040410

Lowry, P. B., Dinev, T., & Willison, R. (2017). Why security and privacy research lies at the centre of the information systems (IS) artefact: Proposing a bold research agenda. *European Journal of Information Systems*, *26*, pp. 546–563. https://doi.org/10.1057/s41303-017-0066-x

Lu, J. (2016). Will Medical Technology Deskill Doctors? *International Education Studies*, *9*(7), pp. 130.

Luo, H., Kou, Z., Zhao, F., & Cai, H. (2019). Comparative life cycle assessment of station-based and dock-less bike sharing systems. *Resources, Conservation and Recycling*, *146*, pp. 180–189. https://doi.org/10.1016/j.resconrec.2019.03.003

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), pp. 305–325. https://doi.org/10.1016/0304-4076(85)90158-7

Malcolm, D. G., Roseboom, J. H., Clark, C. E., & Fazar, W. (1959). Application of a Technique for Research and Development Program Evaluation. *Operations Research*, *7*(5), pp. 646–669. https://www.jstor.org/stable/167013

Mallari, K., Inkpen, K., Johns, P., Tan, S., Ramesh, D., & Kamar, E. (2020). Do i look like a criminal? Examining how race presentation impacts human judgement of recidivism. In R. Bernhaupt (Ed.), *Proceedings of the 2020 CHI conference on human factors in computing systems*. Published by *Association for Computing Machinery*. https://doi.org/10.1145/3313831.3376257

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372867

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, *45*, pp. 135–155. https://doi.org/10.1007/s11747-016-0495-4

Martinez, L. M., Caetano, L., Eiró, T., & Cruz, F. (2012). An optimisation algorithm to establish the location of stations of a mixed fleet biking system: An application to the city of lisbon. *Procedia-Social and Behavioral Sciences*, *54*, pp. 513–524.

Martín, C., Quintana, D., & Isasi, P. (2019). Evolution of trading strategies with flexible structures: A configuration comparison. *Neurocomput.*, *331*(C), pp. 242–262. https://doi.org/10.1016/j.neucom.2018.11.062

Maruping, L., Yin, D., Chen, A., Kankanhalli, A., Burton-Jones, A., & Brown, S. (2025). Quantitative Behavioral IS Research – A Look Back and a Look Forward. *MIS Quarterly*, *49*(1), pp. iii–xviii. https://doi.org/10.25300/MISQ/2025/49.1.00

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant Perspectives During Selection: A Review Addressing "So What?," "What's New?," and "Where to Next?" *Journal of Management*, *43*(6), pp. 1693–1725. https://doi.org/10.1177/0149206316681846

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *42*(2), pp. 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

Megahed, F. M., Chen, Y.-J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2023). How Generative AI models such as ChatGPT can be (Mis)Used in SPC Practice, Education, and Research? An Exploratory Study. *Quality Engineering*, pp. 1–29. https://arxiv.org/abs/2302.10916

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), pp. 1–35. https://doi.org/10.1145/3457607

Mete, S., Cil, Z. A., & Özceylan, E. (2018). Location and coverage analysis of bike-sharing stations in university campus. *Business Systems Research Journal*, *9*(2), pp. 80–95.

Miettinen, K., & Mäkelä, M. M. (2002). On scalarizing functions in multiobjective optimization. *OR Spectrum*, *24*(2), pp. 193–213. https://doi.org/10.1007/s00291-001-0092-9

Milne, G. R., & Rohm, A. J. (2000). Consumer privacy and name removal across direct marketing channels: Exploring opt-in and opt-out alternatives. *Journal of Public Policy & Marketing*, *19*(2), pp. 238–249.

Misra, K., Schwartz, E. M., & Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, *38*(2), pp. 226–252. https://doi.org/10.1287/mksc.2018.1129

Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, *22*(1), pp. 39–72.

Mivule, K., & Turner, C. (2013). A comparative analysis of data privacy and utility parameter adjustment, using machine learning classification as a gauge. *Procedia Computer Science*, *20*, pp. 414–419. https://doi.org/10.1016/j.procs.2013.09.295

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), pp. 529–533. https://doi.org/10.1038/nature14236

Mohammad, Y., Hoda, H., & Andreas, K. (2019). A human-in-the-loop framework to construct context-dependent mathematical formulations of fairness. *AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021)*.

Mostaghim, S., & Teich, J. (2003). Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO). *Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03 (Cat. No. 03EX706)*, pp. 26–33. https://doi.org/10.1109/SIS.2003.1202243

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *The American Economic Review*, *107*(5), pp. 476–480. https://doi.org/10.1257/aer.p20171084

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately.

Journal of Applied Research in Memory and Cognition, *3*(3), pp. 222–229. https://doi.org/10.1016/j.jarmac.2014.05.001
Nair, B. B., Mohandas, V. P., Nayanar, N., Teja, E. S. R., Vigneshwari, S., & Teja, K. V. N. S. (2015). A stock trading recommender system based on temporal association rule mining. *Sage Open*, *5*(2), pp. 2158244015579941. https://doi.org/10.1177/2158244015579941
Niederman, F., & March, S. T. (2012). Design science and the accumulation of knowledge in the information systems discipline. *ACM Transactions on Management Information Systems*, *3*(1), pp. 1–15. https://doi.org/10.1145/2151163.2151164
Ochmann, J., Zilker, S., Michels, L., Tiefenbeck, V., & Laumer, S. (2020). The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations. *ICIS 2020 Proceedings*. International conference on information systems.
Oliver, J. R. (1996). A machine-learning approach to automated negotiation and prospects for electronic commerce. *Journal of Management Information Systems*, *13*(3), pp. 83–112. https://doi.org/10.1080/07421222.1996.11518135
Önüt, S., Kara, S. S., & Işik, E. (2009). Long term supplier selection using a combined fuzzy MCDM approach: A case study for a telecommunication company. *Expert Systems with Applications*, *36*, pp. 3887–3895. https://doi.org/10.1016/j.eswa.2008.02.045
Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine Learning in Information Systems Research. *Management Information Systems Quarterly*, *46*(1), pp. iii–xix. https://aisel.aisnet.org/misq/vol46/iss1/4
Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), pp. 101885. https://doi.org/10.1016/j.jsis.2024.101885
Park, C., & Sohn, S. Y. (2017). An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of seoul. *Transportation Research Part A: Policy and Practice*, *105*, pp. 154–166. https://doi.org/10.1016/j.tra.2017.08.019
Parkes, S. D., Marsden, G., Shaheen, S. A., & Cohen, A. P. (2013). Understanding the diffusion of public bikesharing systems: Evidence from europe and north america. *Journal of Transport Geography*, *31*, pp. 94–103. https://doi.org/10.1016/j.jtrangeo.2013.06.003
Phelps, J., Nowak, G., & Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing*, *19*(1), pp. 27–41.
Pouillot, R., & Delignette-Muller, M. L. (2010). Evaluating variability and uncertainty separately in microbial quantitative risk assessment using two R packages. *International Journal of Food Microbiology*, *142*(3), pp. 330–340. https://doi.org/10.1016/j.ijfoodmicro.2010.07.011
Pournader, M., Ghaderi, H., Hassanzadegan, A., & Fahimnia, B. (2021). Artificial intelligence applications in supply chain management. *International Journal of Production Economics*, *241*, pp. 108250. https://doi.org/10.1016/j.ijpe.2021.108250
Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Savelka, J. (2023). The Robots Are Here: Navigating the Generative AI Revolution in Computing Education. *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*, pp. 108–159. https://doi.org/10.1145/3623762.3633499 pp.108–159
Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., & Miklau, G. (2020). *Fair decision making using privacy-protected data*. pp. 189–199.
Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: *Hv*-block cross-validation. *Journal of Econometrics*, *99*(1), pp. 39–61. https://doi.org/10.1016/S0304-4076(00)00030-0
Rai, A., Constantinides, P., & Sarker, S. (2019). Editor'S comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Q.*, *43*(1), pp. iii–x.
Rastegarpanah, B., Crovella, M., & Gummadi, K. P. (2020). *Fair inputs and fair outputs: The incompatibility of fairness in privacy and accuracy*. pp. 260–267. https://doi.org/10.1145/3386392.3399568
Renkl, A., Atkinson, R. K., & Große, C. S. (2004). How Fading Worked Solution Steps Works – A Cognitive Load Perspective. *Instructional Science*, *32*(1), pp. 59–82.
Rhein-Main-Verkehrsverbund. (24.02.2019). *RMV open data: Infrastructure data: List of RMV-stations*. opendata.rmv.de

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why should i trust you?': Explaining the predictions of any classifier*. pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Rixey, R. A. (2013). Station-level forecasting of bikesharing ridership: Station network effects in three US systems. *Transportation Research Record*, *2387*(1), pp. 46–55.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), pp. 913–929. https://doi.org/10.1111/ecog.02881

Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management*, *57*(3), pp. 739–750. https://doi.org/10.1002/hrm.21852

Rocha, M., & Neves, J. (1999). Preventing premature convergence to local optima in genetic algorithms via random offspring generation. In I. Imam, Y. Kodratoff, A. El-Dessouki, & M. Ali (Eds.), *Multiple approaches to intelligent systems* (pp. 127–136). Published by *Springer Berlin Heidelberg*. pp.127–136

Romero, J. P., Ibeas, A., Moura, J. L., Benavente, J., & Alonso, B. (2012). A Simulation-optimization Approach to Design Efficient Systems of Bike-sharing. *Procedia - Social and Behavioral Sciences*, *54*, pp. 646–655. https://doi.org/10.1016/j.sbspro.2012.09.782

Romero, J. P., Moura, J. L., Ibeas, A., & Benavente, J. (2012). *Car-bicycle combined model for planning bicycle sharing systems*.

Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H.-C. (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics*, *10*(21). https://doi.org/10.3390/electronics10212717

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

Rudloff, C., & Lackner, B. (2014). Modeling demand for bikesharing systems: Neighboring stations as source for demand and reason for structural breaks. *Transportation Research Record: Journal of the Transportation Research Board*, *2430*(1), pp. 1–11. https://doi.org/10.3141/2430-01

Ruß, G., & Brenning, A. (2010). Data mining in precision agriculture: Management of spatial information: 13th international conference on information processing and management of uncertainty in knowledge-based systems, IPMU 2010, dortmund, germany, june 28 - july 2, 2010 : proceedings. In E. Hüllermeier, R. Kruse, & Hoffmann (Eds.), *International conference on information processing and management of uncertainty in knowledge-based systems* (1st ed., pp. 350–395). Published by *Springer Berlin Heidelberg*. https://doi.org/10.1007/978-3-642-14049-5_36

Rybarczyk, G., & Wu, C. (2010). Bicycle facility planning using GIS and multi-criteria decision analysis. *Applied Geography*, *30*(2), pp. 282–293. https://doi.org/10.1016/j.apgeog.2009.08.005

Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L., & Tschantz, M. C. (2020). Measuring non-expert comprehension of machine learning fairness metrics. *Proceedings of the 37th International Conference on Machine Learning*, *119*, pp. 8377–8387. pp.8377–8387

Salas, J., & González-Zelaya, V. (2020). Fair-MDAV: An algorithm for fair privacy by microaggregation. **Modeling Decisions for Artificial Intelligence*: *17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2, 2020, Proceedings*, v. Torra, y. Narukawa, j. Nin and n. Agell (Eds.), Springer International Publishing*, pp. 286–297. https://doi.org/10.1007/978-3-030-57524-3_24

Sambasivan, N., & Veeraraghavan, R. (2022). The Deskilling of Domain Expertise in AI Development. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. https://doi.org/10.1145/3491102.3517578

Samtani, S., Zhu, H., Padmanabhan, B., Chai, Y., Chen, H., & Nunamaker, J. F. (2023). Deep Learning for Information Systems Research. *Journal of Management Information Systems*, *40*(1), pp. 271–301. https://doi.org/10.1080/07421222.2023.2172772

Sarker, S., Chatterjee, S., Xiao Xiao, & 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. https://doi.org/10.25300/MISQ/2019/13747

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, *283*, pp. 103238. https://doi.org/10.1016/j.artint.2020.103238

Sayarshad, H., Tavassoli, S., & Zhao, F. (2012). A multi-periodic optimization formulation for bike planning and bike utilization. *Applied Mathematical Modelling*, *36*(10), pp. 4944–4951.

Schneckenreither, M., & Haeussler, S. (2019). Reinforcement Learning Methods for Operations Research Applications: The Order Release Problem. In G. Nicosia, P. Pardalos, G. Giuffrida, R. Umeton, & V. Sciacca (Eds.), *Machine Learning, Optimization, and Data Science* (pp. 545–559). Published by *Springer International Publishing*. https://doi.org/10.1007/978-3-030-13709-0_46 pp.545–559

Schneider, T., Janson, A., & Schöbel, S. (2018). Understanding the Effects of Gamified Feedback in Mobile Learning – An Experimental Investigation. *ICIS 2018 Proceedings*.

Schöffer, J., Machowski, Y., & Kühl, N. (2021). A study on fairness and trust perceptions in automated decision making. *Joint Proceedings of the ACM IUI 2021 Workshops*.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, pp. 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Schuijbroek, J., Hampshire, R. C., & van Hoeve, W.-J. (2017). Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, *257*(3), pp. 992–1004. https://doi.org/10.1016/j.ejor.2016.08.029

Schworm, S., & Renkl, A. (2007). Learning Argumentation Skills Through the Use of Prompts for Self-Explaining Examples. *Journal of Educational Psychology*, *99*, pp. 285–296. https://doi.org/10.1037/0022-0663.99.2.285

Scott, J., Ho, W., Dey, P. K., & Talluri, S. (2015). A decision support system for supplier selection and order allocation in stochastic, multi-stakeholder and multi-criteria environments. *International Journal of Production Economics*, *166*, pp. 226–237. https://doi.org/10.1016/j.ijpe.2014.11.008

Sengewald, J., & Lackes, R. (2021). The impact of the 'right to be forgotten' on algorithmic fairness. In R. A. Buchmann, A. Polini, B. Johansson, & D. Karagiannis (Eds.), *Perspectives in business informatics research* (pp. 204–218). Published by *Springer International Publishing*. https://doi.org/10.1007/978-3-030-87205-2_14

Sengewald, J., & Lackes, R. (2022). Prescriptive analytics in procurement: Reducing process costs. *Wirtschaftsinformatik 2022 Proceedings*. https://aisel.aisnet.org/wi2022/business_analytics/business_analytics/5

Shen, H., Jin, H., Cabrera, Á. A., Perer, A., Zhu, H., & Hong, J. I. (2020). Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), pp. Article 153.

Shendryk, V., Bychko, D., Parfenenko, Y., Boiko, O., & Ivashova, N. (2019). Information system for selection the optimal goods supplier. *Procedia Computer Science*, *149*, pp. 57–64. https://doi.org/10.1016/j.procs.2019.01.107

Shi, D., Fan, W., Xiao, Y., Lin, T., & Xing, C. (2020). Intelligent scheduling of discrete automated production line via deep reinforcement learning. *International Journal of Production Research*, *58*(11), pp. 3362–3380. https://doi.org/10.1080/00207543.2020.1717008

Shi, J., Guo, J., & Fung, R. Y. K. (2017). Decision support system for purchasing management of seasonal products: A capital-constrained retailer perspective. *Expert Systems with Applications*, *80*, pp. 171–182. https://doi.org/10.1016/j.eswa.2017.03.032

Shokri, R., Strobel, M., & Zick, Y. (2021). *On the privacy risks of model explanations*. pp. 231–241. https://doi.org/10.1145/3461702.3462533

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *5/22/2017*, pp. 3–18. https://doi.org/10.1109/SP.2017.41

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*, pp. 74–87. https://doi.org/10.4018/JDM.2020040105

Skulmowski, A., & Xu, K. M. (2022). Understanding Cognitive Load in Digital and Online Learning: A New Perspective on Extraneous Cognitive Load. *Educational Psychology Review*, *34*(1), pp. 171–196. https://doi.org/https://doi.org/10.1007/s10648-021-09624-7

Linkedin Talent Solutions (2018). *Global recruiting trends 2018. The 4 ideas changing how you hire.* Published by *https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/linkedin-global-recruiting-trends-2018-en-us.pdf*.

Soria-Comas, J., & Domingo-Ferrer, J. (2018). Differentially private data publishing via optimal univariate microaggregation and record perturbation. *Knowledge-Based Systems*, *153*, pp. 78–90. https://doi.org/10.1016/j.knosys.2018.04.027

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Potential for discrimination in online targeted advertising. **Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Sorelle a. Friedler and Christo Wilson (Eds.), New York, NY, USA: PMLR*, pp. 5–19. https://proceedings.mlr.press/v81/speicher18a.html pp.5–19

Spreitzenbarth, J. M., Bode, C., & Stuckenschmidt, H. (2024). Artificial intelligence and machine learning in purchasing and supply management: A mixed-methods review of the state-of-the-art in literature and practice. *Journal of Purchasing and Supply Management*, *30*(1), pp. 100896. https://doi.org/https://doi.org/10.1016/j.pursup.2024.100896

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2459–2468. https://doi.org/10.1145/3292500.3330664 pp.2459–2468

Stadt Frankfurt am Main. (2018-2020). *Offenedaten.frankfurt.de*. www.offenedaten.frankfurt.de

Stadt Frankfurt am Main. (2014). *Frankfurter stadtteilgrenzen für GIS-systeme*. http://offenedaten.frankfurt.de/dataset/d4aaf62a-acd9-4032-bec2-892f4b7e8e7b/resource/842dd252-ad7f-46f2-a064-17f44399dad2/download/stadtteile.zip

Stahmann, P. (2024). AI-based Real-time Anomaly Detection in Digitized Production: Characterization and Socio-technical Facilitation of Decision-making (Doctoral dissertation, Universität Osnabrück).

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing – beyond principles for machine learning. *Journal of Business Research*, *124*, pp. 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Statistischen Amts Stuttgart. (2020). *Kommunis: Informationssystem des statistischen amts*. https://www.domino1.stuttgart.de/web/komunis/komunissde.nsf

Statistisches Amt für Hamburg und Schleswig Holstein. (2018). *Bevölkerung in hamburg am 31.12.2017*.

Suen, H.-Y., Chen, M. Y.-C., & Lu, S.-H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, pp. 93–101. https://doi.org/10.1016/j.chb.2019.04.012

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd ed* (pp. xxii, 526). Published by *The MIT Press*.

Sweeney, L. (1998a). Datafly: A system for providing anonymity in medical data. In T. Y. Lin & S. Qian (Eds.), *Database security XI: Status and prospects* (pp. 356–381). Published by *Boston, MA*. https://doi.org/10.1007/978-0-387-35285-5_22

Sweeney, L. (1998b). *Datafly: A system for providing anonymity in medical data*. pp. 356–381. https://doi.org/10.1007/978-0-387-35285-5_22

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), pp. 571–588. https://doi.org/10.1142/s021848850200165x

Sweeney, L. (2002b). *K*-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), pp. 557–570. https://doi.org/10.1142/S0218488502001648

Sweeney, Latanya. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *ACM Queue*, *11*(3), pp. 10–29. https://doi.org/10.1145/2460276.2460278

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, *22*(2), pp. 123–138. https://doi.org/10.1007/s10648-010-9128-5

Telford, R. J., & Birks, H. J. B. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, *28*(13–14), pp. 1309–1316. https://doi.org/10.1016/j.quascirev.2008.12.020

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the r package sdcMicro. *Journal of Statistical Software*, *67*. https://doi.org/10.18637/jss.v067.i04

Tipping, M. E. (2004). Bayesian inference: An introduction to principles and practice in machine learning. In *Advanced lectures on machine learning* (pp. 41–62). Published by *Springer Berlin Heidelberg*. https://doi.org/10.1007/978-3-540-28650-9_3

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*(sup1), pp. 234–240. https://doi.org/10.2307/143141

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*, pp. 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Trachsel, M., & Telford, R. J. (2016). Technical note: Estimating unbiased transfer-function performances in spatially structured environments. *Clim. Past*, *12*(5), pp. 1215–1223. https://doi.org/10.5194/cp-12-1215-2016

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *Proceedings of the 25th USENIX Conference on Security Symposium*, pp. 601–618. https://dl.acm.org/doi/10.5555/3241094.3241142 pp.601–618

Tran, T. D., & Ovtracht, N. (2018). Promoting sustainable mobility by modelling bike sharing usage in lyon. *IOP Conference Series: Earth and Environmental Science*, *143*, pp. 012070. https://doi.org/10.1088/1755-1315/143/1/012070 pp.012070

Tran, T. D., Ovtracht, N., & d'Arcier, B. F. (2015). Modeling Bike Sharing System using Built Environment Factors. *Procedia CIRP*, *30*, pp. 293–298. https://doi.org/10.1016/j.procir.2015.02.156

United Kingdom. (2018). *Data protection act 2018*. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review*, *22*(3), pp. 271–296.

van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: Support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science*, *43*(5), pp. 615–641.

van den Broek, E., Sergeeva, A., & Huysman, M. (2019). Hiring Algorithms: An Ethnography of Fairness in Practice. *ICIS 2019 Proceedings*, *6*, pp. 6. https://aisel.aisnet.org/icis2019/future_of_work/future_work/6 pp.6

van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2017). Responsible Data Science. *Business & Information Systems Engineering*, *59*(5), pp. 311–313. https://doi.org/10.1007/s12599-017-0487-z

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, pp. 215–222. https://doi.org/10.1016/j.chb.2018.09.009

Vanderschueren, T., Verdonck, T., Baesens, B., & Verbeke, W. (2022). Predict-then-optimize or predict-and-optimize? An empirical evaluation of cost-sensitive learning strategies. *Inf. Sci.*, *594*. https://doi.org/10.1016/j.ins.2022.02.021

Veale, M., Kleek, M. van, & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. https://doi.org/10.1145/3173574.3174014 pp.1–14

Verkehrs- und Tarifverbund Stuttgart GmbH. (25.11.2019). *Vvs open data portal: haltestellen*. www.openvvs.de

Villaronga, E. F., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law &Amp; Security Review*, *34*(2), pp. 304–313. https://doi.org/10.1016/j.clsr.2017.08.007

Wambsganss, T., Kueng, T., Soellner, M., & Leimeister, J. M. (2021). ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. Art. Nr. 683. https://doi.org/10.1145/3411764.3445781

Wang, J., & Lindsey, G. (2019). Do new bike share stations increase member use: A quasi-experimental study. *Transportation Research Part A: Policy and Practice*, *121*, pp. 1–11. https://doi.org/10.1016/j.tra.2019.01.004

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making. In R. Bernhaupt (Ed.), **Proceedings of the 2020 CHI conference on human factors in computing systems*, r. Bernhaupt (ed.), honolulu HI USA, new york, NY, united states: Association for computing machinery* (pp. 1–14). Published by *Association for Computing Machinery*. https://doi.org/10.1145/3313831.3376813 pp.1–14

Wang, X., Lindsey, G., Schoner, J. E., & Harrison, A. (2015). Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development*, *142*(1), pp. 04015001. https://doi.org/10.1061/(asce)up.1943-5444.0000273

Weber, F., Wambsganss, T., Rüttimann, D., & Söllner, M. (2021). Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education. In *ICIS 2021 Proceedings*.

Wei, C.-P., Chen, L.-C., Chen, H.-Y., & Yang, C.-S. (2013.). Mining Suppliers from Online News Documents. *PACIS 2013 Proceedings*. PACIS 2013. https://aisel.aisnet.org/pacis2013/261 Article No.261

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., & Wilson, J. (2019). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1. https://doi.org/10.1109/TVCG.2019.2934619

Wirth, R. and Hipp, J. (2000) CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, 11-13 April 2000, 29-40.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. pp.1–14. https://doi.org/10.1145/3313831.3376781

Winkler, R., Söllner, M., & Leimeister, J. M. (2021). Enhancing problem-solving skills with smart personal assistant technology. *Computers & Education*, *165*, pp. 104148. https://doi.org/10.1016/j.compedu.2021.104148

Wissuchek, C., & Zschech, P. (2024). Prescriptive Analytics Systems Revised: A Systematic Literature Review from an Information Systems Perspective. *Information Systems and e-Business Management*. https://doi.org/10.1007/s10257-024-00688-w

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, *4*.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In R. Mandryk & M. Hancock (Eds.), *Proceedings of the 2018 CHI conference on human factors in computing systems*. Published by *Association for Computing Machinery*. https://doi.org/10.1145/3173574.3174230

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. *Expert Systems with Applications*, *36*(5), pp. 9105–9112. https://doi.org/10.1016/j.eswa.2008.12.039

Wuerzer, T., & Mason, S. G. (2016). Retail gravitation and economic impact: A market-driven analytical framework for bike-share station location analysis in the united states. *International Journal of Sustainable Transportation*, *10*(3), pp. 247–259. https://doi.org/10.1080/15568318.2014.897403

Xu, D., Du, W., & Wu, X. (2021). *Removing disparate impact on model accuracy in differentially private stochastic gradient descent*. pp. 1924–1932. https://doi.org/10.1145/3447548.3467268

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using conditional GAN*. pp. 7335–7345. https://dl.acm.org/doi/10.5555/3454287.3454946

Yang, T.-H., Lin, J.-R., & Chang, Y.-C. (2010). *Strategic design of public bicycle sharing systems incorporating with bicycle stocks considerations*. pp. 1–6. https://doi.org/10.1109/ICCIE.2010.5668312

Zeileis, A., Köll, S., & Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in *R*. *Journal of Statistical Software*, *95*(1). https://doi.org/10.18637/jss.v095.i01

Zhang, M., & Gable, G. (2014). Rethinking the Value of Simulation Methods in the Information Systems Research Field: A Call for Reconstructing Contribution for a Broader Audience. *Research Methods*.

Zhang, Y., Thomas, T., Brussel, M., & van Maarseveen, M. (2017). Exploring the impact of built environment factors on the use of public bikes at bike stations: Case study in zhongshan, china. *Journal of Transport Geography*, *58*, pp. 59–70. https://doi.org/10.1016/j.jtrangeo.2016.11.014

Zhao, J., Deng, W., & Song, Y. (2014). Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in china. *Transport Policy*, *35*, pp. 253–264. https://doi.org/10.1016/j.tranpol.2014.06.008

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., & Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Trans. Evol. Comput.*, *7*. https://doi.org/10.1109/TEVC.2003.810758