On the Interplay of Transparency and Fairness in AI-Informed Decision-Making

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Jakob Schöffer

Tag der mündlichen Prüfung: 11. Oktober 2023

Referent: Prof. Dr. Gerhard Satzger

Korreferent: Prof. Dr. Ali Sunyaev

Abstract

Using artificial intelligence (AI) systems for informing high-stakes decisions has become increasingly pervasive in a variety of domains, including but not limited to hiring, lending, or law enforcement. As the nature of such AI-informed decisions is becoming ever more consequential, numerous examples have shown the potential for adverse effects on humans due to the deployment and operation of such systems. Consequently, researchers and policymakers have articulated the need for heightened scrutiny and initiatives that promote greater transparency and equity in the design and implementation of AI systems.

The first contribution of this thesis is to synthesize the research discourse by conducting a structured literature review. The primary focus is to investigate the intricate relationship between transparency and fairness in AI systems. From this review, we infer several desiderata for transparency, such as its potential to enable fairness audits. However, many of these desiderata lack theoretical or empirical support, which casts doubt on the commonly conjectured role of transparency as a panacea in AI-informed decision-making. These insights form an important cornerstone for all subsequent work presented in this thesis.

The remainder of this thesis studies ramifications of transparency mechanisms as they pertain to various dimensions of fairness. First, we address *fully automated decision-making*. Herein, we propose a ranking-based algorithm that is inherently transparent and respects established notions of fairness. Recognizing the complexity of fairness as a concept, we also assess human perceptions towards AI systems through two mixed-method online studies. In the first study, we observe that perceptions are positively associated with the amount of information that is provided about a given system, ceteris paribus. This could be exploited, however, when malevolent system designers seek to manipulate more vulnerable stakeholders into trusting problematic systems. In the second study, we contrast human perceptions towards AI-based versus human-based decisions. Here, we find that perceptions tend to be more favorable towards automation, although for reasons that may not always be justified.

We then turn to *human-in-the-loop decision-making*, that is, scenarios where humans are endowed with discretionary power to override AI-issued decisions. We first

establish a framework on the interplay of reliance behavior—how humans adhere to or override AI systems—and decision quality. This framework eventually serves as a blueprint for our third online study. The findings from this study indicate that human reliance behavior varies based on whether transparency mechanisms disclose the use of sensitive information, such as gender or race, by the AI system. Interestingly, these differences in reliance behavior imply opposing downstream effects on the fairness of decisions.

Overall, this thesis underscores the complex, multidimensional relationship between transparency and fairness in AI systems. In this respect, it sheds light on the often-overlooked limitations of popular transparency mechanisms and emphasizes the discrepancy between desiderata and empirical evidence. By advocating for a reevaluation of transparency as a more comprehensive concept rather than a monolithic notion, our findings provide valuable insights for researchers and system designers aiming to create genuinely responsible AI systems.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisor, Gerhard Satzger, for his guidance, support, and for fostering a truly unique research environment filled with some of the most inspiring people I have ever met. I am equally thankful to Niklas Kühl for his unwavering support throughout my journey as a PhD student and beyond. I have been learning a lot from you, and I am truly excited about our continued collaboration. I also want to express my gratitude to Ali Sunyaev, Sanja Lazarova-Molnar, and Stefan Nickel for agreeing to be part of my thesis committee.

Research is undeniably a team effort, and I am indebted to my brilliant collaborators, many of whom have become good friends. My deepest gratitude goes to my long-term mentor, Maria De-Arteaga, as well as to Johannes Jakubik, Yvette Machowski, Luca Deck, Michael Vössing, and Isabel Valera. Additionally, I want to express my thanks to my friends from MD4SG/EAAMO: Jessie Finocchiaro, Faidra Monachou, Alex Ritchie, Keziah Naggita, and Marc Juarez. I am genuinely proud of our achievements despite working from different ends of the globe.

Needless to say, I cannot thank my fellow graduate students enough for the wonderful times we shared. In particular, I am grateful to my "roomies" at office 4B-06: Joshua Holstein and Philipp Spitzer, along with former (part-time) residents Lucas Baier, Max Schemmer, and Jannis Walk. A special shoutout goes to Patrick Hemmer, who is not only a talented researcher but also the most loyal gym buddy—our nightly weightlifting sessions were crucial for maintaining my sanity between deadlines.

I am also thankful for various international research travels, where I had the chance to meet old and new friends. Bahar Sarrafzadeh, thank you for an incredible summer internship. Amir Sabet Sarvestani, thanks for being a fantastic mentor. Christian Haas, your hospitality in Vienna made my stay unforgettable, and I look forward to our paths crossing again very soon. Andrew Bell, here is to hoping our collaboration plans become a reality soon. Thank you also to everyone I am forgetting right now.

I am immensely grateful to Alexandra for her invaluable help in reviewing my thesis and, more importantly, her unconditional support and ability to lift my spirits whenever I need it. This means the world to me. Finally, and very importantly, I want to thank my parents, who I do not express gratitude to often enough. Your unending support has made all of this possible. This thesis is dedicated to you.

Contents

Acronyms

AAAI Association for the Advancement of Artificial Intelligence.

ACM Association for Computing Machinery.

ADS Automated Decision System.

AI Artificial Intelligence.

AILIT AI Literacy.

AMTIN Amount of Information.

AVE Average Variance Extracted.

CA Cronbach's Alpha.

CFI Comparative Fit Index.

COMPAS Correctional Offender Management Profiling for Alternative Sanctions.

CR Composite Reliability.

EU European Union.

FAccT Fairness, Accountability, and Transparency.

FTA Fairness Through Awareness.

FTU Fairness Through Unawareness.

GDPR General Data Protection Regulation.

GETA General Equal Treatment Act.

GRE Graduate Record Examination.

HCI Human-Computer Interaction.

IEEE Institute of Electrical and Electronics Engineers.

INFF Informational Fairness.

LIME Local Interpretable Model-Agnostic Explanations.

M Mean.

ML Machine Learning.

MPP Man Professor Predicted as Professor.

MTP Man Teacher Predicted as Professor.

MTT Man Teacher Predicted as Teacher.

NLP Natural Language Processing.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

RMSEA Root Mean Square Error of Approximation.

SD Standard Deviation.

SE Standard Error.

SEM Structural Equation Model.

SHAP Shapley Additive Explanations.

SRCC Spearman's Rank Correlation Coefficient.

SRMR Standardized Root Mean Square Residual.

TLI Tucker-Lewis Index.

TRST Trustworthiness.

ULS Unweighted Least Squares.

US United States.

VIF Variance Inflation Factor.

WPP Woman Professor Predicted as Professor.

WPT Woman Professor Predicted as Teacher.

WTT Woman Teacher Predicted as Teacher.

XAI Explainable Artificial Intelligence.

Part I

Fundamentals

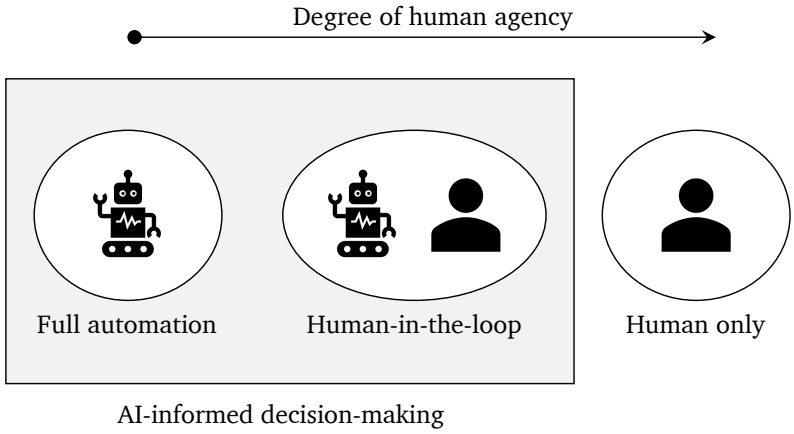
1 Introduction

> *Technology is neither good nor bad; nor is it neutral.*
>
> — **Melvin Kranzberg**
> (American historian)

In recent years, artificial intelligence (AI) has experienced remarkable advancements, profoundly impacting various domains and revolutionizing the way we approach problem-solving and decision-making. AI systems typically harness vast amounts of data and advanced algorithms to analyze complex patterns and provide valuable insights. The utilization of AI in decision-making has been embraced across a multitude of domains, including hiring (Kuncel et al., 2014), lending (Townson, 2020), law enforcement (Heaven, 2020), grading (Satariano, 2020), and healthcare (Leibig et al., 2022), among others. In hiring, for instance, recent studies show that more than 55% of human resource leaders in the United States (US) use predictive algorithms to support hiring activities (Reicin, 2021). Underlying motives of adopting AI systems for informing decisions are manifold, ranging from cost-cutting to improving performance, and enabling more robust and objective decisions (Harris & Davenport, 2005; Kuncel et al., 2014; Newell & Marabelli, 2015).

The degree of AI integration in decision-making processes may vary depending on the specific context. While many tasks may be well-suited for *full automation* through AI systems, others call for greater human oversight. Particularly in certain high-stakes domains, such as those mentioned previously, AI systems often serve as decision support tools that aid human experts, who ultimately bear responsibility for making final decisions. We refer to such settings as *human-in-the-loop* decision-making. For instance, in healthcare, AI systems can play a vital role in assisting clinicians with diagnoses or prognoses. Subsequently, the human experts can utilize these insights to determine the most appropriate course of treatment. In 2022, a collaborative team of cancer researchers from Germany and the US demonstrated the potential of AI systems in the realm of breast cancer screening. Their findings indicated that the combined use of an AI system and a radiologist resulted in superior performance compared to either the standalone AI system or the radiologist working

Fig. 1.1.: Spectrum of AI integration in decision-making processes.



Note: Drawing inspiration from Lai and Tan (2019), we define a spectrum of AI integration in decision-making processes, spanning from full automation *(where an AI system independently issues decisions) to* human only*, in which no AI input is taken into account. Additionally, we introduce* human-in-the-loop *configurations, where humans possess discretionary authority to override initial AI-generated decision recommendations. Finally, we subsume full automation and human-in-the-loop setups under the term* AI-informed decision-making*.*

independently. Similarly, in the realm of criminal justice, a judge might rely on AI-based risk assessment tools when determining bail, ensuring that technology complements human expertise rather than replacing it. We summarize our taxonomy of AI-informed decision-making in Figure 1.1, and will refer back to it as needed. In any event, as AI continues to play an increasingly significant role in shaping decisions, it is crucial to understand and address concerns related to fairness, accountability, and transparency, ensuring that AI-informed decisions align with societal values and ethical standards.

1.1 Motivation

In 2016, a team of investigative journalists from ProPublica sparked a pioneering debate that highlighted concerns regarding the use of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm in the US criminal justice system (Angwin et al., 2016). COMPAS, a risk assessment tool designed to predict the likelihood of an individual reoffending, has been employed by judges in determining bail, sentencing, and parole decisions (Räz, 2022). The tool is property of Equivant/Northpointe (Canton, OH, USA) and based on regression models that rely on features like age, prior arrests, employment status, criminal

history, and other alleged predictors of recidivism or pretrial failure (Northpointe, 2015).

ProPublica's analysis revealed potential biases in the COMPAS algorithm, particularly with regard to racial disparities. The investigation found that Black defendants were more likely to be assigned higher risk scores compared to White defendants, even if they did not ultimately reoffend. Conversely, White defendants were more likely than Black defendants to be incorrectly flagged as low risk (Larson et al., 2016). In technical terms, this means that COMPAS exhibits higher false positive rates and lower false negative rates for Black people compared to White people (Chouldechova, 2017), ultimately putting a higher burden on Black people. This revelation sparked a broader debate around the ethical implications of using algorithms—particularly AI systems—in consequential decision-making processes, which persists to this day. The COMPAS debate also constitutes a significant impetus for this thesis because it surfaced two fundamental concerns regarding the development and implementation of AI systems for decision-making, which have subsequently become the focus of extensive research: fairness and transparency.

1.1.1 Problems Around Fairness

The first concern pertains to the definition of fairness in relation to AI systems. ProPublica's analysis revealed significant discrepancies in AI error rates between demographic groups, leading them to conclude that COMPAS was unfair. Equivant/Northpointe, however, dispute the allegations of racial bias, arguing that the COMPAS system adheres to both calibration (Flores et al., 2016) and predictive parity (Dieterich et al., 2016), which represent alternative conceptions of algorithmic fairness. Interestingly, Chouldechova (2017) and Kleinberg et al. (2017) independently show that these different notions of fairness cannot all be simultaneously satisfied except (i) we have a perfectly accurate predictor, or (ii) the base rates of recidivism (i.e., the proportion of individuals who reoffend in a given population) are equal between demographic groups. Note that both (i) and (ii) are very restrictive conditions that we generally cannot assume to hold in practice.

While Bell, Bynum, et al. (2023) show that we might still hope to construct AI systems that satisfy seemingly incompatible notions of fairness if we allow for some margin of error, these impossibility theorems have far-reaching consequences: they show that algorithmic fairness is not a monolithic concept and cannot possibly be quantified by one-size-fits-all statistical metrics (Bell, Bynum, et al., 2023; Chouldechova, 2017; Friedler et al., 2016). Instead, fairness is a social and ethical concept (Chouldechova,

2017) and intricately dependent on the specific context in which an AI system is deployed (Bell, Bynum, et al., 2023). Although the concept of fairness has recently garnered significant attention in relation to AI systems, it is by no means a new topic. In fact, discussions surrounding fairness, such as the distinctions between equality and equity, have been prevalent in the social sciences and humanities for decades (Cook & Hegtvedt, 1983; Culyer & Wagstaff, 1993).

1.1.2 Problems Around Transparency

The second important theme of the COMPAS debate is with respect to transparency of AI systems. Given incompatible notions of fairness and the seemingly gridlocked debate regarding potential unfairness of the COMPAS algorithm, Rudin et al. (2020), for instance, argue that the focus on fairness might be misguided altogether. Rudin et al. (2020) contend that the debate should instead concentrate on transparency, which they assert is a readily attainable objective that would ultimately enable stakeholders to scrutinize the internal mechanisms of the COMPAS model. They argue that Equivant/Northpointe's utilization of an overly large and complex model with more than 100 input features, coupled with the non-disclosure of their proprietary tool's precise inner workings, has fostered erroneous assumptions about COMPAS, leading to flawed conclusions. Indeed, ProPublica's post-hoc audit of COMPAS was heavily critized for making allegedly wrong assumptions (Chouldechova, 2017; Rudin et al., 2020), including concerns about the inadequacy of the dataset that was used to reverse-engineer the working mechanisms of the tool (Bao et al., 2021; Flores et al., 2016). Rudin et al. (2020) ultimately assert that transparency is intrinsically linked to fairness, as it facilitates a proper vetting of AI systems and, hence, helps surfacing potential issues around fairness.

The COMPAS debate is not the only case where AI systems have been criticized for being biased and opaque, and, as a consequence, adversely affecting individuals or entire demographic groups. Similar cases include an AI recruitment tool used by Amazon that was scrapped after being found to be gender biased, systematically disadvantaging women candidates for technical job roles (Dastin, 2018). The tool was designed to analyze resumes and identify top candidates for various positions within the company. However, it was found to prioritize men applicants, largely due to the man-dominated nature of the tech industry and the historical data the algorithm was trained on. As a result, the tool perpetuated existing gender imbalances, leading Amazon to eventually abandon the project. In 2019, Apple Card's credit limit algorithm came under scrutiny for potential gender bias in its decision-making process after several users had reported that women were offered

significantly lower credit limits compared to men with similar credit profiles (Kelion, 2019). Similar to the COMPAS case, analyses are complicated by the fact that the Apple Card algorithm is proprietary and owned by Goldman Sachs, even though Apple has since published certain information on how the algorithm works (Apple, 2022). For up-to-date information on AI system failures and their consequences, we refer to the AI Incident Database (McGregor, 2021) and the AI, Algorithmic and Automation Incident and Controversy Repository (Pownall, 2019).

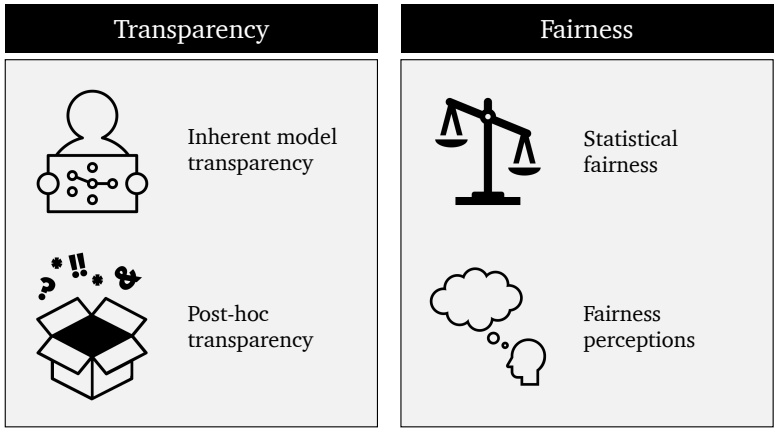
1.1.3 Towards Responsible AI

Based on the aforementioned and similar incidents, researchers, practitioners, and policymakers have called for heightened scrutiny and initiatives that promote transparency and equity in the design and deployment of AI systems in high-stakes decision-making. This has led to the emergence of new research communities, most notably *Fairness, Accountability, and Transparency* (FAccT) and *Explainable Artificial Intelligence* (XAI). Since the COMPAS debate, the FAccT community has put forth a plethora of approaches to operationalize and measure fairness. These concepts can be broadly classified into *statistical fairness notions* (Barocas et al., 2019), such as the ones discussed for the COMPAS debate, and human *fairness perceptions* (Starke et al., 2022). Similarly, the XAI community has proposed numerous *transparency mechanisms* for AI systems. The following definition is a synthesis of different definitions from the literature, primarily from Arrieta et al. (2020) and Lai et al. (2021).

> **Definition 1.1** (Transparency mechanism)**.** *Transparency mechanisms are efforts to make an AI system's functioning or its outputs understandable to relevant human stakeholders.*

Such transparency mechanisms can aim at *inherent model transparency* (Molnar, 2020; Rudin, 2019) or *post-hoc transparency*. The former embodies the idea of using methods that allow relevant stakeholders to see and understand how inputs are mathematically mapped to outputs (Adadi & Berrada, 2018). This typically encompasses rule-based approaches, sparse linear and logistic regressions, as well as tree-based methods (Molnar, 2020). Post-hoc transparency includes popular techniques like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), which can be applied to highly complex models such as deep neural networks. For more information on these techniques, please refer to Chapter 2. The taxonomy of

Fig. 1.2.: Different types of transparency and fairness in AI systems.



transparency and fairness, concisely illustrated in Figure 1.2, serves as a foundational reference throughout this thesis.

Corresponding initiatives in industry are often termed *responsible*, *trustworthy*, or *ethical AI*. IBM, for instance, specifies explainability, fairness, robustness, transparency, and privacy as their foundational pillars of AI ethics (IBM, 2023), and has released several software toolkits to advance these goals, such as `AI Explainability 360` (Arya et al., 2021) and `AI Fairness 360` (Bellamy et al., 2019). Similar programs and tools have been initiated by other companies, for instance, by Microsoft (`InterpretML` (Nori et al., 2019) and `Fairlearn` (Bird et al., 2020)) and Google (Google, 2023), all including transparency and fairness as central elements of responsible AI. For a holistic analysis of trustworthy AI, we refer to Thiebes et al. (2021).

Finally, with the responsible development and deployment of AI systems becoming a pressing global issue, several significant laws and regulations have been introduced to ensure their transparency and fairness. The European Union (EU) has taken the lead in this area, establishing the General Data Protection Regulation (GDPR) in 2018, which sets strict guidelines for personal data processing and transparency. Under GDPR, individuals have the right to know how AI systems are making decisions that affect them, with companies required to provide clear explanations of their AI algorithms. This involves the often-cited "right to explanation" (Goodman & Flaxman, 2017), which requires the disclosure of "the existence of automated decision-making, including [...] meaningful information about the logic involved [...]" to affected individuals (European Union, 2016, Section 2, Article 13). In 2021, the EU proposed the Artificial Intelligence Act (AI Act), further strengthening regulations around AI's ethical use (Madiega, 2021). The AI Act classifies AI systems

based on risk, with high-risk systems facing strict requirements for transparency, human oversight, and unbiased data usage. Both the GDPR and the AI Act, along with similar international laws and regulations, aim to create a framework that fosters trust in AI, ensuring that these technologies are deployed responsibly and with respect for human rights and democratic values.

1.1.4 Challenges Around Operationalization

Despite all endeavors, operationalizing fairness and transparency for AI systems in practice presents several challenges that stem from the inherent complexity and multifaceted nature of these concepts (Arrieta et al., 2020; Friedler et al., 2016; Mulligan et al., 2019).

Firstly, defining fairness in the context of AI can be subjective, as different stakeholders may hold varying perspectives on what constitutes fair treatment. This has been exemplified by the debate surrounding the COMPAS system (Räz, 2022). Reconciling these diverse viewpoints and embedding them in AI systems is a significant challenge and, in many cases, impossible (Chouldechova, 2017; Kleinberg et al., 2017). As indicated by Caton and Haas (2020), the academic literature has put forth a myriad of statistical fairness definitions, each exhibiting unique advantages and disadvantages, as well as a strong dependency on context, goals, and ethical considerations surrounding a given AI system (Bell, Bynum, et al., 2023). Worryingly, Corbett-Davies and Goel (2018) demonstrate that a considerable number of these prevalent fairness metrics exhibit notable statistical shortcomings. In response, a growing number of researchers have adopted a more human-centered approach, focusing on fairness perceptions (Starke et al., 2022). This approach aims to elucidate human attitudes towards AI systems in general or, more specifically, towards individual AI-informed decisions. However, prior research has shown that human perceptions are brittle (Grgić-Hlača et al., 2020; Harrison et al., 2020; Nyarko et al., 2021) and easily misled—particularly by transparency mechanisms (Chromik et al., 2019; Eiband et al., 2019; Lakkaraju & Bastani, 2020). Moreover, it remains often unclear how human perceptions inform behavior. For instance, it appears important to examine how high or low trust and fairness perceptions of decision subjects (i.e., individuals affected by AI-informed decisions) relate to their propensity to appeal AI-informed decisions. Similarly, we might be interested in understanding how human-in-the-loop decision makers' perceptions towards an AI system relate to their reliance on its recommendations. Despite the significance of these inquiries for the development of effective AI systems, these questions remain largely unanswered.

Secondly, the academic community has proposed an extensive array of transparency mechanisms, and it often remains unclear in practice which are most suitable. Existing laws and regulations remain vague as well: Goodman and Flaxman (2017, p. 55), referring to the GDPR requirements, suppose that "any adequate explanation would, at a minimum, provide an account of how input features relate to predictions." But such *feature-based explanations*, while popular in both research and practice (Bhatt et al., 2020; Gilpin et al., 2018), come with their own set of challenges—which we discuss in more detail in Chapter 2.

Models that are inherently transparent are frequently considered the gold standard in terms of rendering AI systems comprehensible to pertinent stakeholders (Candelon et al., 2023; Rudin, 2019). However, their simplicity by design may preclude high predictive performance (Gunning & Aha, 2019). Moreover, while it is commonly assumed that their inner workings are readily understandable, it is far from obvious that lay people can accurately interpret the implications of, for instance, feature coefficients in regression models. From a practical standpoint, it is also unlikely that organizations will employ inherently transparent models because they cannot make profits from intellectual property afforded to more sophisticated models like deep neural networks (Rudin, 2019). For these and other reasons, it has become common practice to draw upon post-hoc transparency techniques, which are typically used to explain *black-box models*, that is, models that do not fall into the category of being inherently transparent. Such transparency mechanisms are often claimed to fulfill a variety of desiderata for different stakeholders, many of which relate to fair and ethical decision-making (Langer, Oster, et al., 2021; Lipton, 2018). For instance, in a recent Forbes article (Kite-Powell, 2022, p. 1), it is claimed that "companies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums." Other desirable effects have been suggested, among others, by Dodge et al. (2019, p. 275), claiming that transparency mechanisms "provide a more effective interface for the human in-the-loop, enabling people to identify and address fairness and other issues." However, it appears that empirical evidence has yet to substantiate many of these claims—a sentiment echoed recently by Balkir et al. (2022) and Langer, Oster, et al. (2021), among others.

1.2 Research Design

The concepts of transparency and fairness, including their interaction, are nuanced and complex. Despite the evident overarching need for both, numerous ambiguities

persist regarding their practical operationalization. Through this thesis, we offer a comprehensive analysis of this intricate relationship between transparency and fairness, advancing the broader academic endeavor of understanding and improving AI-informed decision-making as a whole. We make algorithmic, conceptual, and empirical contributions, and we place a particular focus on human-centered evaluations, assessing the impact of popular transparency mechanisms on both statistical and perceived fairness in AI-informed decision-making.

Many claims have been made on what transparency mechanisms are to provide with respect to fairness in AI-informed decision-making. A natural first endeavor of this thesis is to subsume these desiderata, as they have been formulated in the scientific literature.

Research Question RQ1
What are the desiderata for transparency mechanisms with respect to fairness in AI-informed decision-making?

We answer **RQ1** by conducting a structured review of the relevant literature since 2016, the year of the COMPAS debate and, as a consequence, the genesis of most pertinent scientific work. We analyze a total of 169 research papers to infer a set of nine canonical claims that are commonly being made on the assumed effects of transparency mechanisms on fairness in AI-informed decision-making. More generally, we find that the role of transparency is (i) multifaceted, with different stakeholders pursuing varying fairness desiderata, (ii) intricate due to the lack of definitive evidence supporting many of these desiderata, and (iii) multimodal because transparency mechanisms can affect fairness in diverse ways. We also encounter multiple assertions suggesting transparency and fairness as conflicting objectives. A notable example is provided by Kleinberg and Mullainathan (2019), who argue that simple (i.e., inherently transparent) models incentivize the use of sensitive demographic information to the disadvantage of marginalized groups. Based on these theoretical insights, we are interested in how such trade-offs can be dealt with in practice.

Research Question RQ2
How can we design AI systems that are inherently transparent and fair?

To answer **RQ2**, we instantiate an AI system based on a novel ranking-based algorithm. Our algorithm ranks instances based on predefined monotonic relationships between individual input features and the outcome of interest, thereby respecting an important criterion for inherently transparent AI systems (Molnar, 2020). The algorithm also mutes the influence of sensitive information, such as gender or age, on the final decision. We show theoretically that our method is consistent with the notion of *fairness through awareness*, which deems an AI system fair if similar individuals are treated similarly (Dwork et al., 2012). Finally, we demonstrate empirically that our algorithm yields favorable results with respect to balanced outcome distributions between men and women compared to traditional supervised machine learning. This distinction is particularly notable in the presence of strong label bias in observational training data—for instance, when men were disproportionately favored in the past. Overall, our study illustrates that AI systems can be designed to simultaneously achieve inherent transparency and popular notions of statistical fairness. However, fairness is not a purely technical concept; hence, we are also interested to learn how humans perceive such systems.

Research Question RQ3
How do transparency mechanisms affect people's fairness perceptions towards AI systems?

We address **RQ3** through a mixed-method behavioral study. More concretely, we conduct a randomized between-subjects experiment with lay people to assess how humans perceive an AI-based lending system when they are provided with different types of information about said system. We find that study participants perceive the AI system as more informationally fair when they receive more information about it. Informational fairness is a construct that subsumes different facets of adequateness of a system's explanations regarding its inner workings and outcomes. We also find that high perceptions of informational fairness translate to high perceptions of the system's reliability and integrity. This raises concerns as it suggests that system designers might coax less powerful stakeholders into trusting AI systems merely by inundating them with information. In the same study, we solicit study participants' input on several open-ended questions about their perceptions, from which we infer their desiderata for explanations. Among others, they suggest that explanations should (i) be consistent, that is, not convey conflicting information; (ii) convey precisely how a given feature value impacts the prediction; and (iii) be actionable, empowering stakeholders to achieve their respective goals.

Given the prevalent demand for human oversight in AI-informed decision-making, we conduct a follow-up mixed-method study to examine how people's perceptions may change when the final decision is made by a human versus an AI system, ceteris paribus.

Research Question RQ4
How do people's fairness perceptions differ towards a human versus an AI system as the final decision maker?

As a response to **RQ4**, we see, perhaps surprisingly, that study participants perceive the human decision maker as *less* fair, and the difference in perceptions becomes even more significant for participants with high self-reported AI literacy. From the qualitative analysis of open-ended responses, we furthermore conclude that the preference for full automation often hinges on the misguided assumption that data-driven decision-making is objective and, hence, inherently fair.

Recognizing that human perceptions towards AI systems are often miscalibrated, we turn to measuring how transparency mechanisms affect human *behavior* and downstream metrics related to fairness and accuracy of decisions. To this end, we study human-in-the-loop decision-making setups, where human experts retain the discretionary authority to override AI recommendations. Herein, prior work has argued that transparency is an essential mechanism to enable human decision makers to make better and fairer decisions (Arrieta et al., 2020; Dodge et al., 2019; Gilpin et al., 2018). To properly study such claims, however, we need to holistically understand the interplay between human reliance behavior (i.e, when and how human decision makers accept or override AI recommendations) and decision quality.

Research Question RQ5
What is the relationship between human reliance on AI-based decision recommendations and common measures of decision quality?

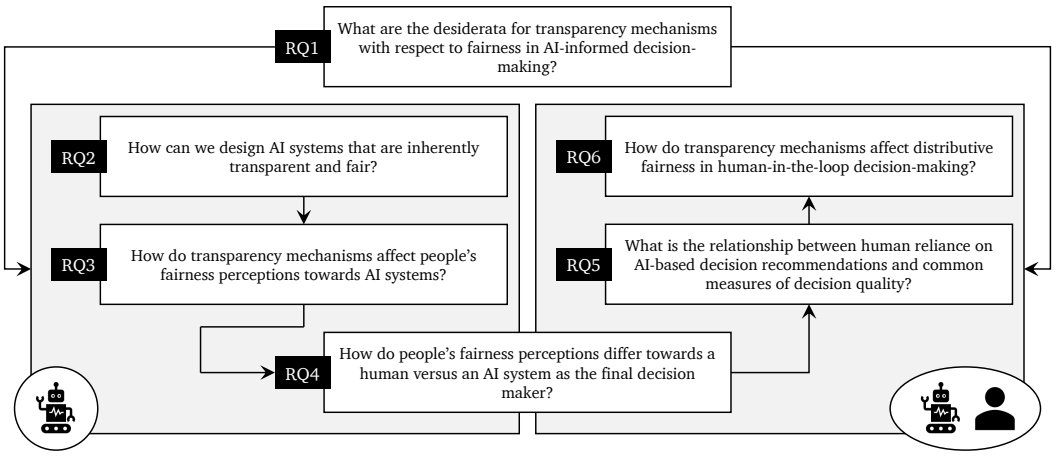
We address **RQ5** by proposing a comprehensive theoretical framework that formalizes the interplay between reliance and decision quality in human-in-the-loop systems. We advocate that in order to fully grasp the effects of transparency mechanisms on decision quality, it is crucial to analyze the mediating role that human reliance on individual AI recommendations plays. The importance lies in discerning

how mechanisms affect human propensity to accept or override correct and wrong AI recommendations, and, in turn, how this behavioral response influences measures of decision quality, such as accuracy and fairness. Our proposed framework goes one step further: it disentangles the effects of transparency mechanisms on both the *quantity* and *quality* of reliance; that is, the frequency at which humans adhere to AI recommendations versus their ability to accurately assess the correctness of AI recommendations. This theoretical construct lays the groundwork for our following empirical study, where we examine the influence of transparency mechanisms on humans-in-the-loop regarding their capacity to enhance the fairness of decisions. This study also connects our previous findings by exploring the relationship between fairness perceptions and statistical notions of fairness.

Research Question RQ6
How do transparency mechanisms affect distributive fairness in human-in-the-loop decision-making?

We address **RQ6** by conducting a randomized between-subjects experiment. In this study, we examine the effects of popular feature-based explanations on individuals' ability to augment distributive fairness—defined as the absence of disparities in types of erroneous decisions across genders—for an occupation prediction task. We focus particularly on how any effects are mediated by human fairness perceptions and reliance on AI recommendations. Crucially, we find that feature-based explanations do not enable study participants to differentiate between correct and wrong AI recommendations. Instead, we show that they may affect reliance behavior irrespective of the correctness of AI recommendations. Depending on which features an explanation indicates to be considered by the AI system, this can foster *or* hinder distributive fairness: when explanations highlight features that are task-irrelevant and evidently associated with sensitive information (e.g., on gender), this prompts overrides that *counter* stereotype-aligned AI recommendations. Meanwhile, if explanations appear task-relevant, this induces reliance behavior that *reinforces* stereotype-aligned errors. These results show that feature-based explanations are not a reliable mechanism to improve distributive fairness, as it has been shown that the use or disuse of sensitive information is neither a necessary nor a sufficient condition for distributive fairness (Apfelbaum et al., 2010; Kleinberg et al., 2018; Nyarko et al., 2021). Figure 1.3 on page 15 summarizes all research questions that this thesis addresses, including their logical flow counterclockwise from top left to top right.

Fig. 1.3.: Summary of research questions addressed in this thesis.



Note: Arrows indicate the logical flow; ***RQ1*** *informs both our work on fully automated (left) and human-in-the-loop (right) decision-making.*

1.3 Thesis Structure

This thesis is structured in four parts. Part I covers general foundations, Part II focuses on transparency and fairness in fully automated scenarios, Part III explicitly looks at human-in-the-loop decision-making, and Part IV concludes this work. The general structure of this thesis is outlined in Figure 1.4 on page 16, including references to published and unpublished articles that parts of the thesis are based upon. The thesis is written in a self-contained manner where each chapter can be read and understood on its own, independent of the rest. Concurrently, we aim to avoid excessive repetitiveness for a more engaging reading experience.

In Part I, we lay the foundations for all subsequent studies. The present Chapter 1 motivates this work and outlines its research design. Chapter 2 gives a concise overview of background information that is of relevance for the remainder of this thesis. This includes information on transparency and fairness, as well as on human agency and stakeholders in AI-informed decision-making. Finally, in Chapter 3, we present the findings from our structured review and qualitative analysis of the pertinent literature, answering **RQ1**.

In Part II, we study AI systems that may be leveraged for fully automated decision-making. First, in Chapter 4, we develop a ranking-based system to showcase how AI systems can be built to respect both inherent transparency and popular notions of statistical fairness—which addresses **RQ2**. We then empirically assess human perceptions of fairness towards AI system in the presence of transparency

Fig. 1.4.: Structure of this thesis.

Part I: Fundamentals

1. Introduction

2. Background

3. Fairness Desiderata of Transparency

Schöffer, J., Deck, L., De-Arteaga, M. & Kühl, N. (2023). Overcoming intuitions: A critical survey on fairness benefits of explanations. *Working Paper.*

Part II: Fully Automated Decision-Making

4. Designing Inherently Transparent and Fair AI Systems

Schöffer, J., Kühl, N. & Valera, I. (2021). A ranking approach to fair classification. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '21).* **Acceptance rate: 37%**.

5. Assessing Fairness Perceptions Towards AI Systems

Schöffer, J., Kühl, N. & Machowski, I. (2022). "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. *ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).* **Acceptance rate: 25%**.

6. Comparing Fairness Perceptions Towards AI-Based Versus Human-Based Decisions

Schöffer, J., Machowski, I. & Kühl, N. (2022). Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making. *55th Hawaii International Conference on System Sciences 2022 (HICSS-55).* **Acceptance rate: 48%**.

Part III: Human-In-The-Loop Decision-Making

7. Conceptualizing the Interdependence of Reliance Behavior and Decision Quality

Schöffer, J.,* Jakubik, J.,* Vössing, M., Kühl, N. & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. *2nd International Conference on Hybrid Human-Artificial Intelligence (HHAI 2023).* **Acceptance rate: 43% | Best paper award: top 2% of submissions**.

*denotes equal contribution

8. Assessing Transparency Effects on Reliance Behavior and Fairness of Decisions

Schöffer, J., De-Arteaga, M.* & Kühl, N.* (2023). On explanations, fairness, and appropriate reliance in human-AI decision-making. *Under Review. Preliminary Version: ACM CHI 2023 Workshop on Trust and Reliance in AI-Assisted Tasks (TRAIT).* **Acceptance rate: 24% (full presentation)**.

*denotes equal contribution

Part IV: Conclusion

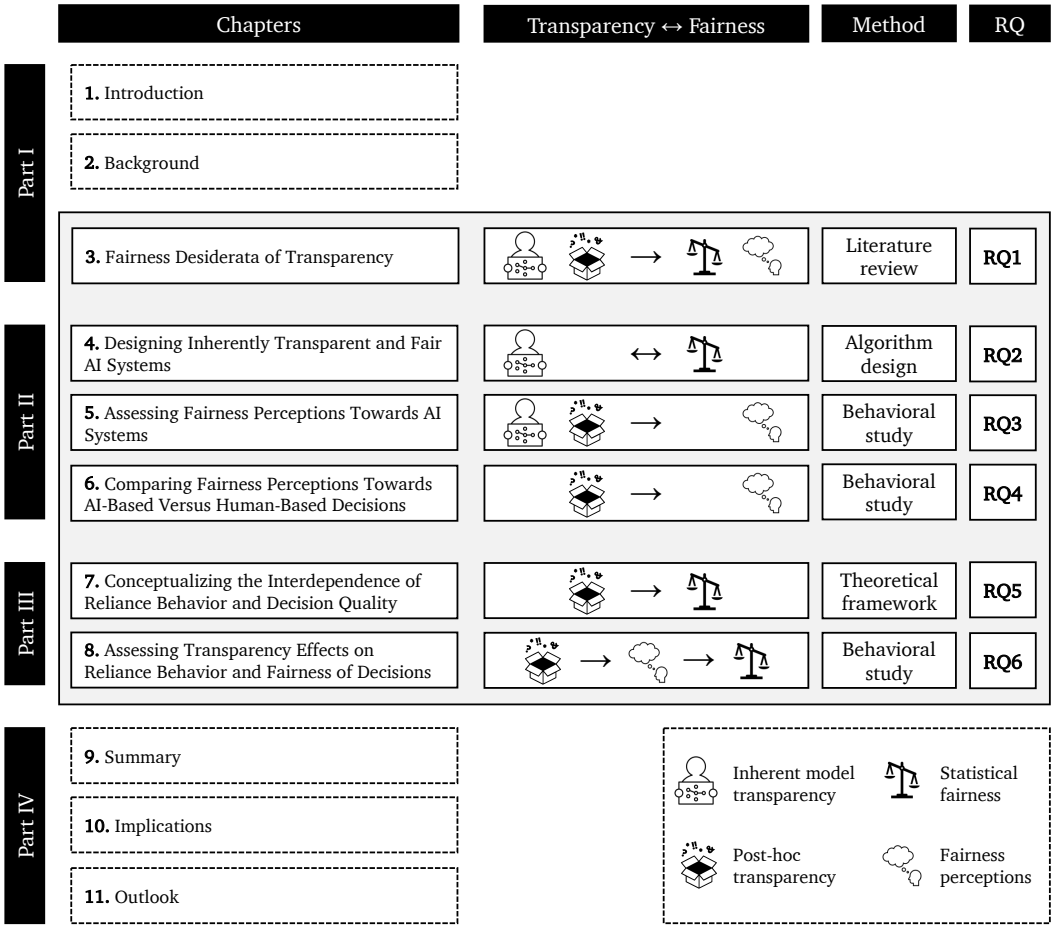
9. Summary

10. Implications

11. Outlook

mechanisms (Chapter 5), and we draw comparisons of perceptions between full automation and scenarios in which the final decision is made by a human (Chapter 6). These two behavioral studies address **RQ3** and **RQ4**, respectively.

Fig. 1.5.: Overview of methods and research questions by thesis chapter.



Chapter 6 serves as a segue into Part III, where we explicitly study human-in-the-loop decision-making.[1] Such settings involve a two-stage decision-making process: an AI system issues an initial decision recommendation, which is then forwarded to a human expert, who accepts or overrides it. The practice of accepting or overriding AI recommendations is commonly referred to as *reliance behavior*. In Chapter 7, we first establish a theoretical framework that elucidates the interdependence of reliance behavior and decision quality—which effectively answers **RQ5**. This framework

[1] Note that the systems studied in Part II could readily be used in human-in-the-loop settings as well. The difference to Part III is that here we explicitly study the role of human decision makers and how they rely on AI-issued recommendations.

ultimately serves as a foundational guide for our following investigation of **RQ6**. Chapter 8 empirically studies the effects of popular feature-based explanations on enabling humans to augment distributive fairness over a baseline setting without explanations. By honing in on the mediating role of reliance behavior and fairness perceptions, this chapter forges links with other sections of this thesis. It builds upon and synthesizes previous findings of this thesis, culminating in a comprehensive analysis of AI-informed decision-making when humans act as the "last line of defense against AI failures" (Passi & Vorvoreanu, 2022, p. 1). Figure 1.5 on page 17 summarizes the structure of this thesis, including the facets of transparency and fairness that are at the center of each analysis, the methods that we apply, as well as the research questions that each chapter addresses.

2 Background

In this chapter, we present a succinct collection of foundations that serve as a valuable resource for the remainder of the thesis. This chapter does not aim to provide an exhaustive review of all related work. Instead, we delve deeper into the intricate relationship between transparency and fairness in Chapter 3. Additionally, each subsequent chapter of this thesis offers more specific details on the relevant background and related scholarly work.

In recent years, artificial intelligence (AI) has rapidly evolved into a transformative force across various domains, reshaping our understanding of information processing and decision-making. The emergence of AI can be traced back to Alan Turing's seminal work on *Computing Machinery and Intelligence* (Turing, 1950). The term *artificial intelligence* itself gained popularity during the *Dartmouth Summer Research Project on Artificial Intelligence*, held in 1956 (McCarthy et al., 2006). Since then, many different definitions of AI have emerged, as summarized by S. J. Russell and Norvig (2021). For the purpose of this thesis, we adopt a slightly modified version of the European Commission's timely and relevant definition from 2018 (Smuha, 2018).

Definition 2.1 (AI system)**.** *AI systems are computer systems designed by humans that, given a complex goal, act by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data, and deciding the best action(s) to take to achieve the given goal.*

When such AI systems are applied in the context of decision-making, we refer to that as *AI-informed decision-making*. In brief, AI-informed decision-making seeks to harness the computational capabilities and data-driven insights of AI systems to improve the efficiency, accuracy, and reliability of decisions across a wide array of disciplines (Colson, 2019). In Section 2.2, we discuss this is more detail, exploring different degrees to which AI systems inform decisions. These degrees include full automation, where AI systems make final decisions, as well as human-in-the-loop decision-making, in which humans retain discretionary power to override AI-issued decision recommendations. The remainder of this chapter is structured as

follows: in Section 2.1, we establish important technical foundations of AI systems. After Section 2.2, where we establish the role of human agency in AI-informed decision-making, we provide a summary of relevant stakeholders including their goals and incentives in Section 2.3. Finally, we provide background information on transparency and fairness in AI systems in Sections 2.4 and 2.5, covering inherent transparency and post-hoc transparency, as well as statistical fairness and fairness perceptions.

2.1 Technical Preliminaries

Throughout this thesis, we utilize notation that is partially borrowed from supervised machine learning (ML). Concretely, we mostly adhere to the notation used by Barocas et al. (2019). However, not all AI systems rely on ML. As Kühl et al. (2022) point out, they may also employ other statistical methods, including deterministic decision rules. Hence, ML is typically considered a subset of AI. This distinction will become important when we discuss transparency of AI systems in Section 2.4, as well as in Chapter 4, where we develop an AI system that does not rely on ML.

According to the previously introduced definition, an AI system utilizes various pieces of observational information, denoted as *features* and represented by X. It then maps X to a plausible decision $\widehat{Y}$, sometimes also called a *prediction*, with respect to a target whose true value Y is unknown at the time of decision-making. In this thesis, we only consider cases where Y and $\widehat{Y}$ take on discrete, mostly binary, values. In such cases, the mapping is also called a *classifier* or *predictor*, and we may refer to it as f. It is important to note that an AI system typically consists of more than just a classifier (M. Lee et al., 2022). While a classifier is its central technical component, an AI system may also include an interface and other elements, possibly with explanations. Yet, certain characteristics of the classifier extend to the overall system. For instance, when we lack knowledge about the specific functional form of the classifier, we commonly refer to the entire system as a *black box*. Similarly, when we assess an AI system's performance or fairness, these are, at least in part, properties of the underlying classifier.

To provide a concrete illustration, consider the case of an AI system that is used for informing financial lending decisions. In this case, a bank aims to identify individuals who are likely to repay their loans while rejecting those who are less likely to do so. However, it is crucial to note that the information regarding who will actually repay their loans, Y, is not available at the time of decision-making. The AI

system under consideration takes as input a set of features X, which may consist of information on an applicant's income, their credit score, employment history, and more. Subsequently, this information is processed and mapped to a decision, denoted as $\widehat{Y} = f(X)$. Ideally, $\widehat{Y}$ serves as a reasonably good estimate of Y, but in general we would expect to have some instances where $\widehat{Y} \neq Y$.

To evaluate the performance of AI systems, a natural way to do so is by computing the share of instances where $\widehat{Y} = Y$. This percentage is called the *accuracy* of a classifier:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}.$$

While accuracy is a widely used measure, we may sometimes be interested in a more granular analysis of an AI system's errors. Let us refer back to the lending example. Consider specifically the following two cases, where we assume $Y, \widehat{Y} \in \{\checkmark, \text{✗}\}$, with ✓ denoting a positive decision to lend, and ✗ a negative decision to deny the loan:

(i) The AI system predicts $\widehat{Y} = \checkmark$, but we have $Y = \text{✗}$.

(ii) The AI system predicts $\widehat{Y} = \text{✗}$, but we have $Y = \checkmark$.

While both of these cases are mistakes, their ramifications are drastically different: in case (i), the AI system incorrectly suggests granting a loan to an applicant who is not creditworthy. Conversely, in case (ii), a loan is denied despite the applicant's ability to repay it.[1] The type of error that occurs in (i) is typically referred to as a *false positive*, whereas the error in (ii) is called *false negative*. Especially when we are concerned with fairness analyses, testing for potential disparities in rates of false positives and false negatives across different demographic groups will be of significant importance. We address this is more detail in Section 2.5.

As noted earlier, AI systems apply techniques such as ML or other statistical methods to arrive at a performant classifier f (Kühl et al., 2022). For instance, f could be derived from a simple logistic regression, a sophisticated deep neural network, or even be a deterministic decision rule. If f is based on ML, it is sometimes also referred to as a *model*. It is worth noting that logistic regression or decision rules are commonly considered inherently transparent models (Molnar, 2020). On the other hand, deep neural networks are often regarded as black-box models, meaning that $f(X) = \widehat{Y}$ involves a complex nonlinear mapping (LeCun et al., 2015), which is not

[1] This assumes that we have access to all true values of Y at the time of evaluation, which may not always be the case in practice. Instead, we may only be able to observe Y later, or not at all. Several solutions to this *selective labels problem* have been proposed in the literature (De-Arteaga et al., 2018; Lakkaraju et al., 2017). For clarity of exposition, however, we still assume that reliable access to Y is possible.

readily comprehensible to all stakeholders. We speak to transparency of AI systems in more depth in Section 2.4.

In practical applications, the majority of AI systems utilize classifiers that rely on supervised ML (Jordan & Mitchell, 2015). Recent advancements in AI have also seen tremendous success by leveraging combinations of different learning paradigms: OpenAI's ChatGPT, for instance, uses (semi-)supervised ML, reinforcement learning, as well as transfer learning to arrive at what has become perhaps the most popular AI tool ever created (K. Hu, 2023). In this thesis, unless explicitly investigating inherent model transparency, we do not make assumptions about how classifiers are initialized or what their specific functional forms are. For most of the thesis, it is adequate to consider f as a given but unknown function that takes a set of features as input and provides a recommended decision (i.e., a prediction) as output.

2.2 Human Agency in AI-Informed Decision-Making

AI systems can be employed in various ways to support decision-making processes. Lai and Tan (2019) distinguish different levels of human agency in these processes, ranging from full automation, where decisions are made without human involvement, to entirely human-made decisions.[2] Advocates of automation have put forth several arguments, including potential cost-cutting benefits and the ability to make more robust, objective, and overall improved decisions (Colson, 2019; Harris & Davenport, 2005; Kuncel et al., 2014; Newell & Marabelli, 2015). Furthermore, the use of AI systems has been seen as a means to counter human stereotyping in areas such as recruitment (Chalfin et al., 2016; Koivunen et al., 2019), healthcare (Grote & Berens, 2020; Triberti et al., 2020), or financial inclusion (Lepri et al., 2017).

2.2.1 Putting a Human in the Loop

As outlined previously, however, AI systems often learn from historical data, which itself may be biased (Mehrabi et al., 2021). In fact, it has been shown that AI systems may not only perpetuate but *exacerbate* existing biases, leading to adverse effects on individuals or entire demographic groups (Eubanks, 2018). An alarming array of unsettling examples illustrate these hazards. In addition to the instances previously discussed in this thesis, they include race and gender stereotyping in

[2] It is crucial to emphasize that when discussing fully automated decision-making, it does *not* imply that such systems lack any human oversight or appropriate safeguards.

job ad delivery (Imana et al., 2021), discrimination of Latinx and African-American borrowers in algorithmic mortgage loan pricing (Bartlett et al., 2022), car insurance pricing systems that disadvantage foreign-born drivers based on birthplace and gender (Fabris et al., 2021), and more. The newly proposed EU AI Act, among other laws and regulations, therefore demands that no AI systems be used when their deployment poses an "unacceptable risk" that clearly threatens people's safety, livelihoods, and rights (Madiega, 2021, p. 5). This prohibition includes AI systems engaged in harmful manipulation, systems that exploit vulnerable groups (e.g., people with disabilities), or those utilized for social scoring. Other systems, particularly those categorized as "high risk"—including systems utilized in hiring practices or law enforcement—must adhere to certain regulations. These encompass critical areas such as "data governance, transparency, [and] human oversight" (Madiega, 2021, p. 6).

Researchers and policymakers have started to argue that high-stakes decision-making warrants heightened involvement of humans. These arguments are supported by two distinct strands of reasoning. The first strand is rooted in moral philosophy and legal studies. It postulates that humans uniquely, not AI systems, possess the capability to exercise discretionary judgment on a case-by-case basis, and that human discretion is an essential component of justice (Binns, 2022). Numerous discussions highlight the important role of human empathy and subjective judgment in decision-making (Decety & Fotopoulou, 2015; Mencl & May, 2009), qualities which remain beyond the grasp of AI systems. Similarly, it has been argued that human discretion is necessary for reasons of liability and responsibility (Bryson et al., 2017; Wagner, 2019), as well as to uphold the rule of law (Hildebrandt, 2018; Zalnieriute et al., 2019). Barocas et al. (2019, pp. 23–43) provide a comprehensive account on the legitimacy of automation and its inherent limitations.

The second line of justification for increased human agency adopts a more pragmatic viewpoint, suggesting at its core that human involvement in the decision-making process can improve decision quality. One prominent argument posits humans as potential safeguards against AI failures, essentially acting as a "last line of defense" (Passi & Vorvoreanu, 2022, p. 1). The related idea of *human-AI collaboration* acknowledges both AI systems and humans as distinct decision makers, each endowed with their own strengths and weaknesses. Although there are numerous ways to operationalize such human-AI collaboration, the overarching objective is to combine the efficiency and scalability of AI systems with the profound contextual understanding inherent to humans. By incorporating a human into the decision-making loop, the intention is to harness the unique strengths of both entities while mitigating their weaknesses. Such *human-AI complementarity* is eventually spec-

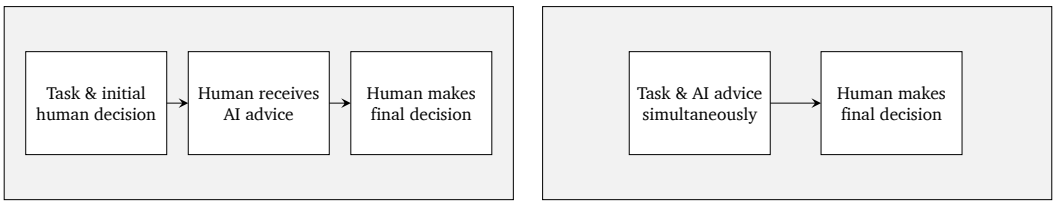
ulated to outperform the decision-making abilities of either the AI system or the human operating in isolation (Hemmer et al., 2021).

2.2.2 Operationalizing Human-In-The-Loop Decision-Making

The integration of humans and AI systems for decision-making has been the subject of extensive research across various domains. These domains encompass a wide range of areas, including industrial settings such as tool wear analysis (Treiss et al., 2021) and smart manufacturing (Garcia et al., 2019), healthcare applications like breast cancer detection (Leibig et al., 2022) and COVID-19 diagnosis (Tsai et al., 2021), as well as law and public services, which involve risk assessment (Green & Chen, 2019a) and child maltreatment hotline screening (De-Arteaga et al., 2020), among many others. For a comprehensive list of domains, we refer to Lai et al. (2021).

Various approaches to integrating humans and AI systems for decision-making have been proposed. Tejeda et al. (2022), for instance, distinguish *sequential* versus *concurrent* setups, as illustrated in Figure 2.1. In the sequential paradigm, the human initially makes a decision based on the given task. Subsequently, that same human is presented with an AI recommendation and is then responsible for making a final call, which may or may not incorporate the AI advice. This setup has been conceptualized and empirically assessed, for instance, by Schemmer et al. (2023). One notable advantage of the sequential setup is its ability to disentangle the effect of AI advice on human decision-making by allowing researchers to observe whether and how the presence of such advice may change the initial human guess.

Fig. 2.1.: Sequential and concurrent paradigms for integrating humans and AI systems for decision-making, as defined by Tejeda et al. (2022).



(a) Sequential paradigm (b) Concurrent paradigm

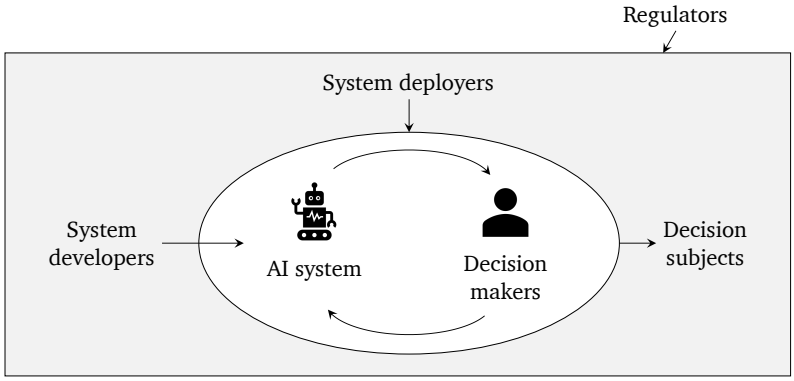
However, as pointed out by Tejeda et al. (2022), many decision-making scenarios in the real world do not include independent human-only decisions prior to AI involvement. Instead, these situations align more closely with the concurrent setup—which we study in Part III of this thesis. In this concurrent paradigm, AI advice

is presented simultaneously with the prediction task, without prior measurement of a human decision. This makes the process simpler and less time consuming, but it also hampers the interpretation of empirical findings because any observed agreement between human and AI does not necessarily indicate that the human actively adopted the AI recommendation. Instead, it would also be possible that the human just ignored the AI input and reached the same decision independently.

Both the sequential and concurrent paradigms, along with their various derivatives, have found application in numerous settings. Nonetheless, empirical studies have yielded inconclusive evidence with respect to human-AI complementarity (Hemmer et al., 2021; Langer, Oster, et al., 2021; Schemmer, Hemmer, Nitsche, et al., 2022). While some make a case for the human-in-the-loop, for instance, in child maltreatment hotline screening (De-Arteaga et al., 2020), others have observed instances where human participation actually leads to a decrease in decision-making performance compared to a fully automated baseline (Green & Chen, 2019a). The latter is particularly likely when the AI system already demonstrates high performance in isolation, as we show in Chapter 7 of this thesis. Instead of humans and AI systems complementing each other, we often see that humans rely too little or too strongly on AI recommendations, regardless of the actual AI performance (J. D. Lee & See, 2004; van Dongen & van Maanen, 2006). These tendencies are often associated with the phenomena of algorithm aversion (Dietvorst et al., 2015) and automation bias (Goddard et al., 2012), respectively—and causes for either are multifaceted (De-Arteaga et al., 2020).

To facilitate human-AI complementarity, a number of different interventions have been proposed, the most prominent of which are transparency mechanisms, which are a major theme of this thesis. For an overview of other assistive interventions in human-in-the-loop decision-making, we refer to the work of Lai et al. (2021). Additionally, novel frameworks to integrating humans and AI systems have emerged, such as learning-to-defer strategies. Here, AI systems have the ability to either make a decision, typically when they are highly confident, or defer to a human expert. Such approaches have been proposed by Hemmer et al. (2023), Madras et al. (2018), and Mozannar and Sontag (2020), among others, and effectively implemented in different domains, such as breast cancer screening (Leibig et al., 2022). Interestingly, Madras et al. (2018) argue that the act of deferring to human experts can positively influence the fairness of decisions, a topic that we will revisit in Chapter 8 of this thesis.

Fig. 2.2.: Stakeholders associated with AI-informed decision-making, as defined by Langer, Oster, et al. (2021).



Note: The human decision maker is only present in human-in-the-loop settings, and missing in fully automated ones. Arrows address the relationships between stakeholders.

2.3 Stakeholders in AI-Informed Decision-Making

AI-informed decision-making involves a diverse range of stakeholders, and it is crucial to identify and understand their roles, particularly in discussions related to transparency and fairness, which are the central themes in this thesis. In this section, we review these stakeholders as they are commonly grouped in the literature, along with their respective action spaces, representative goals, and incentives. Previous research has summarized stakeholders within the context of system integration, encompassing the following categories: system developers, system deployers, decision makers, decision subjects, and regulators (Arrieta et al., 2020; Langer, Oster, et al., 2021; Preece et al., 2018). It is important to note that these categories are not mutually exclusive. Individuals involved in system development can themselves be impacted by the systems they create, and stakeholders across categories may share similar goals. For instance, both decision makers and decision subjects may seek fairness within specific contexts. In the following, we list the stakeholders roughly in order of their occurrence in the AI system lifecycle (De Silva & Alahakoon, 2022). We visually summarize all stakeholders of AI-informed decision-making in Figure 2.2, based on the taxonomy by Langer, Oster, et al. (2021).

2.3.1 System Developers

This group of stakeholders encompasses individuals involved in the development and distribution of AI systems, including product managers, data scientists, ML engineers,

and user experience designers. It would be overly simplistic to assume that all stakeholders within AI system development companies share identical goals. For instance, system designers may prioritize enhancing the user experience by creating intuitive interfaces, particularly for humans-in-the-loop. On the other hand, ML engineers are concerned with debugging models and optimizing system performance (Langer, Oster, et al., 2021). Moreover, these relevant companies often operate as entities with a fiduciary responsibility to their shareholders. Consequently, certain forces within these organizations may seek increased acceptability, engagement, and utilization of their systems by end users (e.g., decision makers). This can result in a desire for certain mechanisms, such as providing explanations of AI systems, to support that objective. These explanations might be used for persuading humans to follow a given decision, elucidating the relevance of product recommendations, and increasing user buy-in. We elaborate more on such *dark patterns* (Chromik et al., 2019) as well as unintended negative downstream consequences of transparency mechanisms in Section 2.4.

2.3.2 System Deployers

The category of AI system deployers encompasses a wide range of individuals and institutions, from educational administrators implementing student success prediction systems to hospital managers integrating decision support tools for doctors, or even governments. While they may not directly use an AI system themselves, they see benefit in what the system offers to their organizations. For this group of stakeholders, deploying AI systems typically involves significant costs but also holds the promise of savings or generating additional revenue. For instance, a university may increase its revenue by improving retention rates and consequently boosting enrollment (Millea et al., 2018). When governments are deploying AI systems, expected benefits may also pertain to fostering social welfare. Considering the potential financial savings and value that AI systems can bring, deployers often have a vested interest in ensuring user acceptance of such systems (Langer, Oster, et al., 2021), which may be facilitated through transparency mechanisms. A second desideratum, according to Langer, Oster, et al. (2021), is legal compliance, because deployers are generally held accountable for unlawful functioning of AI systems.

2.3.3 Decision Makers

Decision makers are the (end-)users of AI systems, whose decisions are the primary target of augmentation through AI-based support. They are typically domain experts in their respective fields, and we often refer to them as *humans-in-the-loop*. Doctors, judges, mortgage lenders, and hiring managers are all decision makers who are typically using such AI systems. In fully automated decision-making, this stakeholder group may not exist. The desires of decision makers are commonly geared towards making better decisions according to some specified criteria. Other desiderata of decision makers are with respect to the user experience of AI systems, including improved usability, education on how these systems ought to be used, and satisfaction with the overall interaction experience (Langer, Oster, et al., 2021). A seemingly overlooked goal of this group is that they need to be armed with enough information to convey a given decision and the reasons behind it to those affected by AI-informed decision. Decision makers often serve as intermediaries between the system itself and the decision subjects. As such, decision makers might need to help clients (e.g., job applicants) understand why a certain decision was made, and enable those adversely affected by an AI system to "challenge its outcome based on plain and easy-to-understand information" (OECD, 2019, p. 8).

2.3.4 Decision Subjects

In this thesis, our focus revolves around decision-making processes that directly impact humans. Consequently, there is always a decision subject involved. Decision subjects are the individuals for whom decisions are being made. This group of stakeholders encompasses various individuals, including loan applicants, medical patients, job seekers, and others. In the context of consequential decisions, decision subjects are often the most vulnerable stakeholders. Unfair, incomprehensible, or otherwise problematic decisions can have far-reaching repercussions for them. Consequently, decision subjects frequently raise important questions regarding the reasoning behind system decisions and the fairness of their treatment. Hence, key concerns for this group of stakeholders include fairness and ethical aspects of the decision-making process and its outcomes, as identified in previous studies (Langer, Oster, et al., 2021). To that end, it has been argued that transparency mechanisms can serve as valuable tools for decision subjects (Arrieta et al., 2020). An example of this is the provision of explanations for credit scores, where individuals can understand the primary factors influencing their scores and may take appropriate actions to improve their creditworthiness for future positive judgments.

2.3.5 Regulators

Regulatory agencies are often government bodies tasked with overseeing the development and deployment of AI systems for consequential decision-making. Examples include the European Commission or the US Food and Drug Administration (FDA). This group of stakeholders is typically interested in ensuring fair and accountable systems with informed consent (Langer, Oster, et al., 2021). Concerns around fairness and transparency have recently become more urgent, with the US Federal Trade Commission (FTC) in 2021 warning that "deception [or] discrimination" resulting from AI systems could constitute law enforcement action (Jillson, 2021, p. 1). Similarly, violations of the proposed EU AI Act will possibly result in penalties of up to 6% of a company's global turnover, or 30 million euros for private entities (Kop, 2021). For more information on algorithmic auditing, we refer to Raji and Buolamwini (2022).

In summary, this section has highlighted the diverse array of stakeholders involved in AI-informed decision-making, each with their unique goals and incentives. Many of these objectives are not compatible, which is a critical factor to consider when determining the intended audience for transparency mechanisms and identifying whose fair treatment we are aiming to ensure.

2.4 Transparency in AI-Informed Decision-Making

Transparency mechanisms have been touted as a panacea in AI-informed decision-making, solving various concerns of different stakeholders, which arise from the opacity of many AI systems (Langer, Oster, et al., 2021). Lazar (2022) outlines three reasons when and why such opacity arises:

(i) AI systems are often *intellectual property* of companies, as we have seen for the COMPAS tool earlier in this thesis.

(ii) AI systems often employ methods, especially ML, which are *too complex* to be fully understood by all stakeholders.

(iii) The use of high-dimensional ML models (e.g., deep learning) in AI systems results in nonlinear feature mappings that often identify *unintuitive correlations*.

The primary aim of transparency mechanisms is to tackle issues related to opacity, thereby enabling a "justified understanding" of AI systems (Lazar, 2022, p. 1). This implies that a mere sense of understanding is insufficient; the understanding must

be grounded in legitimate comprehension. Lazar (2022) further argues that such a justified understanding hinges on the unique goals and capabilities of stakeholders. This underlines the importance of effectively conveying the information that a stakeholder seeks to obtain, and presenting it in a manner that is penetrable for the respective individuals.

Achieving a justified understanding of AI systems may be important for a variety of reasons. As Doshi-Velez and Kim (2017, p. 3) argue, the quest for transparency typically arises from an "incompleteness in the problem formulation." This means that merely acquiring a prediction from an AI system is insufficient for some problems. In such cases, transparency emerges as a vehicle for fulfilling other essential requirements of AI systems. Adadi and Berrada (2018) identify four downstream desiderata that depend on a thorough understanding of AI systems: justification, control, improvement, and discovery. *Justification* pertains to the ability to explain decisions to affected individuals, particularly when unexpected outcomes arise. Adadi and Berrada (2018) posit that transparency mechanisms act as a means to validate AI-informed decisions as fair, ethical, and compliant with laws and regulations. *Control* signifies the notion that an understanding of AI systems can assist in identifying and rectifying potential shortcomings. *Improvement* is predicated on the idea that comprehending why an AI system delivers certain results enables human stakeholders to "make [the system] smarter" (Adadi & Berrada, 2018, p. 52143). Finally, *discovery* denotes the potential for humans to gain new insights through transparency, fostering a transfer of knowledge from the AI system to the human user. A more granular list of desiderata for transparency is given by Langer, Oster, et al. (2021): while transparency is often heralded as a silver bullet in the realm of AI-informed decision-making, they highlight a considerable discrepancy between expectation and reality. Some desiderata lack empirical scrutiny altogether, while other empirical studies regarding the effects of transparency mechanisms yield ambiguous, if not contradictory, results. We provide a comprehensive analysis of transparency desiderata specifically with respect to fairness in Chapter 3.

The scholarly literature differentiates between two kinds of AI system transparency: inherent model transparency and post-hoc transparency, each necessitating distinct techniques. We will briefly explore these aspects of transparency and their corresponding mechanisms in the following sections. We also provide a concise overview of common criticism against transparency mechanisms in Section 2.4.3. Note that our intent is not to deliver an exhaustive review of all existing techniques. For comprehensive insights, we refer to the survey work by Adadi and Berrada (2018), Arrieta et al. (2020), Molnar (2020), Bell, Stoyanovich, and Nov (2023), and others.

2.4.1 Inherent Model Transparency

In essence, mechanisms focused on inherent model transparency seek to address challenges connected to aspects (ii) and (iii) highlighted at the beginning of Section 2.4, by restricting the complexity of a predictor, as highlighted by Molnar (2020). This approach favours the use of relatively simple predictors, enabling relevant human stakeholders to understand the mathematical relationship between input features and outcomes. Additionally, it helps avoid the risk of the model identifying complex and unintuitive correlations. Although the notion of "simplicity" in this context is open to interpretation, and there is no definitive boundary separating inherently transparent models from those classified as black boxes, inherently interpretable models usually adhere to certain properties that make them allegedly easier to understand.

Molnar (2020) outline three typical properties of inherently transparent models. The first one is *linearity*, where associations between features and outcomes are modelled linearly. Secondly, *monotonicity*, which means that an increase or decrease in a feature value always affects the outcome in one particular direction. For instance, when a higher income invariably increases (or at least not decreases) the chances of getting a loan. S. Wang and Gupta (2020) also refer to this as a *deontological* property, because in many cases it may be unethical to use certain information (e.g., years of work experience) as negative evidence. Finally, the third property stipulates that inherently transparent models should not account for "too many or too complex *[feature] interactions*" (Molnar, 2020, p. 37). The inclusion of such interactions could potentially compromise the model's comprehensibility. AI systems that satisfy all three of the above properties are usually based on linear or logistic regression, decision trees, or decision rules (Molnar, 2020).

The appeal of inherently transparent AI systems is undeniable, and numerous scholars and practitioners have advocated for their implementation, for instance, Rudin (2019) and Candelon et al. (2023). However, certain applications necessitate the use of more advanced models capable of identifying complex patterns. In fact, the ability to unearth intricate correlations and learn complex representations is often the primary reason why these systems are utilized in the first place (LeCun et al., 2015). This requirement has been underscored in diverse areas such as natural language processing (Bubeck et al., 2023), computer vision (Redmon et al., 2016), and bioinformatics (Jumper et al., 2021). In these fields, highly nonlinear models with billions of parameters are frequently employed. Moreover, the premise of inherently interpretable models sharply contrasts with the reality that commercial entities typically aim to capitalize on the intellectual property derived from AI

systems. This would be unattainable if their systems were so simple and inherently transparent that anyone could reproduce them. Given this context, our attention now shifts to the concept of post-hoc transparency. This notion revolves around attempts to enhance transparency in black-box models retrospectively, particularly when the inner workings of these models are not readily accessible—either because they are unknown or because they are not understandable by relevant human stakeholders (Guidotti et al., 2018). For instance, a deep learning model used for predicting credit defaults could be open source, but the complex functional description of the classifier has no interpretation that allows humans to readily obtain a justified understanding of the system.

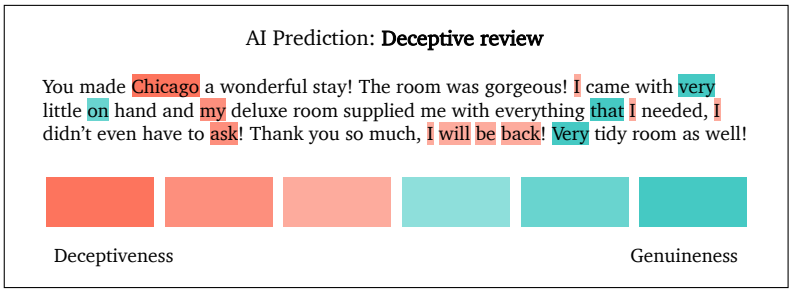
2.4.2 Post-Hoc Transparency

In contrast to inherent model transparency, which is a characteristic of the respective classifier, post-hoc transparency mechanisms seek to make AI systems comprehensible using various techniques layered on top of an already established classifier, often a complex and opaque model based on deep learning. These post-hoc transparency mechanisms are occasionally referred to as *explanations*.[3] While there exists a wide range of *model-specific explanations*, that is, those applicable only to certain ML models, our focus lies solely on *model-agnostic explanations* due to their significance in both research and practical applications (Binns et al., 2018; ElShawi et al., 2021). These model-agnostic mechanisms effectively decouple the explanation from the model itself (Molnar, 2020). Consequently, model-agnostic explanations can be leveraged even with inherently transparent models, serving as an interface to convey a model's internal mechanics to relevant human stakeholders.

Post-hoc transparency mechanisms encompass a range of techniques, and numerous taxonomies have been proposed to structure them. The scientific literature distinguishes techniques that seek to explain individual predictions (*local explanations*) and those aiming to elucidate the overall functionality (*global explanations*) of an AI system (Guidotti et al., 2018). However, it has been argued that an understanding of global model behavior can be achieved by aggregating local explanations (Lundberg et al., 2020). Local model-agnostic explanations, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), have seen rising popularity in literature (Adadi & Berrada, 2018). The general idea of LIME, for instance, is to create a

[3] It is worth noting that *explanations* are sometimes referred to more broadly as any interface between human and AI system, with the aim of making the system comprehensible to the human (Guidotti et al., 2018). In such instances, explanations may pertain to both inherently transparent models and black-box models.

Fig. 2.3.: Exemplary feature-based explanation for deceptive review detection.



Note: Adapted from Lai et al. (2020).

simple auxiliary model that locally approximates the behavior of a given black-box classifier. This surrogate is then used to gauge the contribution of individual features to the prediction of interest. Therefore, they are also referred to as *feature-based explanations*.

Local model-agnostic explanation techniques can generate visual tools such as saliency maps for computer vision tasks (Simonyan et al., 2014) or highlight important words for text classification (Lai et al., 2020). These visual cues indicate whether a particular feature had a positive or negative influence on a specific outcome, and also measure the strength of that contribution. See Figure 2.3 for an exemplary feature-based explanation for the task of deceptive review detection, taken from Lai et al. (2020). However, given that the majority of post-hoc explanations are predicated on approximations of the original black-box model, Rudin (2019) advises caution, stating that such explanations may inevitably misrepresent the original model in certain areas of the feature space. Concerns regarding the faithfulness of explanations to the underlying model have also been raised and discussed by other researchers such as Jacovi and Goldberg (2020) and Slack, Hilgard, et al. (2020).

Additional prevalent types of post-hoc explanations include *counterfactual explanations* (Wachter et al., 2018), which indicate the smallest necessary modification to an input feature that would change the AI system's predicted outcome to a desired value. Another type is *case-based explanations*, where a historical case (along with its corresponding outcome) that is most similar to the one being explained is shown to stakeholders (Binns et al., 2018). The subject of explanations has a rich history in the social sciences and humanities, with discussions centered around the properties that constitute an explanation and the criteria for an effective explanation (Lombrozo, 2012; T. Miller, 2019; Mittelstadt et al., 2019). While delving into the specifics of these works is beyond the scope of this thesis, there is much to be learned from them in the quest for designing effective explanations of AI systems.

2.4.3 Critique of Transparency Mechanisms

Earlier, we highlighted that transparency is typically credited with a plethora of benefits in AI-informed decision-making. However, many desiderata associated with transparency have not been empirically evaluated, and for others, there is inconclusive evidence as to whether existing mechanisms can live up to these hopes (Langer, Oster, et al., 2021). For instance, prior work has shown that transparency mechanisms can impact human perceptions both in positive and negative ways (Starke et al., 2022). Other findings suggest that explanations may or may not be beneficial with respect to decision quality in human-in-the-loop decision-making (Schemmer, Hemmer, Nitsche, et al., 2022). In order to design and deploy effective transparency interventions, it will be crucial to analyze why we observe such opposing effects, and to understand in more detail the mechanisms through which transparency affects relevant desiderata. This thesis contributes to that.

An even more disconcerting issue that has surfaced is the potential for explanations to mislead stakeholders, either deliberately or inadvertently. While system developers typically aim to provide optimal service to their users, there are unfortunate instances where they may resort to deceptive practices or exploit their users. This often involves the use of *dark patterns* (Chromik et al., 2019; A. Narayanan et al., 2020). Explanations, despite their seemingly benign nature, are not exempt from this risk. There is a possibility that developers with malicious intent may use explanations as a tool to persuade users into actions that may not align with their best interests. This is not a mere hypothetical scenario: Lakkaraju and Bastani (2020) construct faithful explanations to deceive people into trusting AI systems that make decisions based on sensitive information (e.g., race or gender) by leveraging correlations between sensitive and non-sensitive features. Similar techniques have been proposed by Pruthi et al. (2020), Dimanov et al. (2020), and others. Chromik et al. (2019) explore multiple other realizations of dark patterns related to transparency mechanisms based on previous explorations of dark patterns in other interfaces. In the case of *placebic explanations*, which are explanations that do not provide any insight into the underlying AI system, Eiband et al. (2019) discover that humans may display a level of trust comparable to that evoked by real explanations. This suggests that merely having explanations can enhance trust in AI systems.

Lima et al. (2022) argue that explanations may also be exploited by system developers as a tactic to sidestep responsibility. This could be done by conferring a misguided sense of power and agency onto more vulnerable stakeholders, such as decision subjects (i.e., humans affected by AI-informed decisions). This concern has been recently echoed by Liao and Wortman Vaughan (2023), specifically in

relation to large language models. Even without any ill intentions, Ehsan and Riedl (2021) underscore several challenges that stem from unanticipated negative effects of explanations. These include misplaced trust in AI or inaccurate estimations of an AI system's capabilities, either overestimating or underestimating its potential. Several other limitations of transparency mechanisms are discussed by Ananny and Crawford (2018). In Chapter 3, we address common critique of transparency mechanisms particularly with respect to fairness.

There are various reasons, including those previously mentioned, that have led certain researchers to assert that transparency is not always necessary or even justified. One such concern is that if humans fully understand which features an AI system considers and the methodology behind them, they may find ways to game the system in order to "unfairly receive goods or services" (Ananny & Crawford, 2018, p. 979). Let us consider credit scoring as an illustrative example: Molnar (2020) elaborates on instances where loan applicants, armed with an understanding of the AI system, might inflate their credit scores without actually enhancing their capacity to pay back loans. For instance, knowing that maintaining more than two credit cards negatively impacts their credit score could lead them to simply return excess credit cards. This, in turn, artificially boosts their credit score, even though the number of credit cards is just an indicator, not a causal determinant, of their creditworthiness. The problem of manipulation that is enabled by transparency has also been discussed by Diakopoulos (2016). Other instances where transparency may not be needed include scenarios where an AI system informs decisions of insignificant impact (e.g., vacation planning or movie recommendations), or whenever a problem is well studied, such as optical character recognition (Molnar, 2020).

2.5 Fairness in AI-Informed Decision-Making

The increasing integration of AI systems in decision-making processes across various sectors brings risks of discrimination and unfairness into focus. In addition to the scenarios elucidated in Chapter 1, numerous instances have been reported where AI systems have contributed to contentious decisions in realms as diverse as online advertising (Datta et al., 2015), health services (Barda et al., 2021), and predictive policing (Lum & Isaac, 2016), among others. Each of these instances employs AI systems to inform consequential decisions which have subsequently been critiqued for "bias," "unfairness," or "discrimination." These terms are frequently confounded in research and practice. Hence, we provide a succinct clarification in the following.

The term *bias* has several interpretations that intersect and sometimes contradict. In statistical contexts including ML, *bias* has a specific meaning revolving around systematic errors. Selection bias, for instance, refers to estimation errors resulting from non-random sampling in a population (Heckman, 1979), for instance, when members of certain demographic groups are more likely to to be sampled than others (Campolo et al., 2017). AI systems trained on such biased data may propagate this bias, manifesting as disparities in predictive accuracy across demographic groups. This segues into the normative interpretation of *bias*, which alludes to judgments based on preconceived notions or prejudices. From this perspective, *bias* is often defined as a "tendency which prevents unprejudiced consideration" (Pannucci & Wilkins, 2010, p. 619). While there may be a tendency to limit discussions of bias to the statistical sense when dealing with AI systems, it is essential to acknowledge that statistical and normative interpretations of biases often intersect in practice: statistically biased models can result in unequal and unfair outcomes for different societal groups (De-Arteaga et al., 2022). While the normative notion of bias may be hard to disentangle from the terms *unfairness* or *discrimination* (Campolo et al., 2017), *unfairness* is generally construed as any adverse impacts on humans that arise from (statistical) biases in AI systems (Mehrabi et al., 2021). Finally, in line with Barocas et al. (2019), we use the terms *unfairness* and *discrimination* roughly as synonyms in the context of this thesis.

In the following, we first establish normative foundations around allegations of unfairness in AI-informed decision-making. This necessitates a brief exploration of how moral and political philosophers have been thinking about justice and fairness in more general terms. We then make the connection to AI systems. Subsequently, we discuss multiple approaches that Fairness, Accountability, and Transparency (FAccT) researchers have employed to measure and combat unfairness within AI systems. We start by covering relevant statistical notions of fairness as they pertain to AI systems in Section 2.5.3. We then introduce and discuss the idea of "fairness through unawareness" (Kusner et al., 2017, p. 2) in Section 2.5.4, which is commonly believed to be an effective pathway to fairness but has severe limitations. Lastly, in Section 2.5.5, we explore human-centered approaches to fairness, specifically those around human perceptions.

2.5.1 Normative Foundations

At the outset of this section, it appears essential to ponder a critical question—what does *fairness* really mean, especially in the context of AI systems? Before we delve into specific notions of fairness that have been adopted in the AI literature, it is

important to examine some normative underpinnings of claims related to bias and unfairness in AI systems. The definition of *fairness* has been a point of intense discussion over several decades, particularly in the fields of moral and political philosophy, as well as economics.

Fairness and Justice The concept of *fairness* is closely related to that of *justice*, and both terms may in some instances be viewed as synonyms, as acknowledged by Velasquez et al. (1990). However, we may think of *justice* as a standard of moral rightness (Goldman & Cropanzano, 2015) and the idea of "giving each person what they deserve" (Velasquez et al., 1990, p. 1). As such, it is a broader concept than fairness. In fact, Goldman and Cropanzano (2015, p. 313) argue that justice "denotes conduct that is morally required, whereas 'fairness' denotes an evaluative judgment as to whether this conduct is morally praiseworthy." As a result, assessments of fairness tend to exhibit more variability within a population compared to assessments of justice (Goldman & Cropanzano, 2015). A central pillar of justice since Aristotle is to "treat like cases alike" (Gosepath, 2021, p. 1) and "dissimilar cases differently, proportionally to their differences" (Goodin, 1999, p. 189). This idea is practically reflected in the notion of *fairness through awareness* (Dwork et al., 2012) in the context of AI systems (see Section 2.5.2). Similarly, justice has also been framed as the "opposite of arbitrariness," meaning that it requires an "impartial and consistent application of rules" (D. Miller, 2021, p. 1). Thus, it can be argued that justice, in its application, is concerned with granting fair treatment (Velasquez et al., 1990). We make every effort to uphold these terminological nuances in our work. However, we acknowledge that the distinction between *justice* and *fairness* has not always been consistently applied in the realm of AI research.

We have seen that the principle of giving everyone what they deserve is central to discussions around justice and fairness, but it remains vague until we specify precisely what people deserve and on what grounds—we address this in more detail shortly. Generally, conflicts around justice and fairness surface in scenarios where individuals lay claim to certain rights, opportunities, or resources that may potentially be at odds with others'. In such cases, justice is invoked as a mechanism to resolve these disputes by delineating what each person is rightfully entitled to (D. Miller, 2021). Conversely, Velasquez et al. (1990, p. 1) note that "there would be no point of talking about justice or fairness if it were not for the conflicts of interest that are created when goods and services are scarce and people differ over who should get what." Typical examples for such scarce resources are certain jobs or money. Importantly, as D. Miller (2021) points out, we cannot label circumstances as unjust if no agent, be it an individual or an institution, has played a role in

their creation. This is relevant in discussions around accountability in AI-informed decision-making—which are, however, mostly beyond the scope of this thesis. In the following, we address relevant differences in scope and operationalization of justice and fairness.

Procedural and Distributive Justice Justice has many facets, two of which are especially relevant for our considerations around AI systems: distributive and procedural justice. According to Lamont and Favor (2017, p. 1), "principles of distributive justice are [...] best thought of as providing moral guidance for the [...] processes and structures that affect the distribution of benefits and burdens in societies." When extrapolated to AI systems, the concept of distributive justice[4] is primarily focused on the outcomes these systems generate. Examples of such outcomes include the acceptance or denial of loan requests or job applications. A just distribution demands that the distributor disseminates available resources based on certain pertinent criteria, such as equity, merit, or need (D. Miller, 2021). Procedural justice, on the other hand, is concerned with the "procedures that might be used to determine how benefits and burdens of various kinds are allocated to people" (D. Miller, 2021, p. 1).

The relationship between procedural and distributive justice is of particular interest for this thesis. Rawls (1999) distinguishes three types of procedural justice: *perfect procedural justice* corresponds to situations where adherence to a specified procedure inevitably leads to a fair division based on some independently defined criterion, although this is seldom achievable in reality. *Imperfect procedural justice*, on the other hand, involves situations where following a just procedure does not necessarily guarantee a fair result, like court trials. Lastly, *pure procedural justice* pertains to cases where there is no independent standard for fairness of the outcome separate from the process. In such scenarios, the act of adhering to a fair procedure and achieving a fair outcome cannot be disentangled—this applies, for instance, to gambling (Rawls, 1999).

Such considerations around procedural justice or fairness are notably applicable to our discussions on AI systems. In this context, FAccT researchers occasionally refer to procedural fairness when they reason about the mapping of input features to outcomes (i.e., the predictor) and how this mapping affects the fairness of predictions. For instance, some researchers have posited that whether an AI system utilizes or disregards sensitive information on individuals (such as their gender or

[4]The FAccT literature also refers to this as *distributive fairness*. Similarly, the terms *procedural justice* and *procedural fairness* are often used interchangeably (Goldman & Cropanzano, 2015).

race) is relevant to its procedural fairness (Grgić-Hlača, Zafar, et al., 2018). However, it is generally acknowledged that this is not a sufficient criterion for fairness of outcomes in general (Kleinberg et al., 2018). Hence, the nuanced taxonomy of Rawls (1999) lets us conclude that withholding sensitive information from an AI system does not entail perfect procedural justice. We address this in more detail in Section 2.5.4.

Egalitarianism Justice is often seen as the universal principle for judging whether an action is ethically acceptable or desirable (D. Miller, 2021). However, the application of this principle can take many forms that vary significantly across time and cultures. As a result, it is inevitable that diverse, sometimes conflicting, interpretations of justice exist. Interestingly, as pointed out by Binns (2018), the contemporary discourse around fairness in AI-informed decision-making often resonates with certain considerations pertaining to the philosophical theory of *egalitarianism*. This idea advocates for equality in a way that "people should get the same, or be treated the same, or be treated as equals, in some respect" (Arneson, 2013, p. 1). An important follow-up question centers on identifying exactly *what* should be equally distributed among individuals. Cohen (1989) also refers to this as the "currency" of egalitarianism, aiming to answer a fundamental question that was raised by Sen (1979): *Equality of What?* As summarized by Binns (2018) and Arneson (2013), egalitarians have proposed a variety of competing answers to this question, which include equality of welfare (e.g., preference-satisfaction), resources (e.g., income), or capabilities, which denote the ability and means to accomplish certain tasks. Adding to this complexity, Walzer (1983) adopts a relativistic perspective, asserting that justice, and especially equality, is contingent on context. In practice, this implies that we "cannot assume that some fairness metrics which are appropriate in one context will be appropriate in another" (Binns, 2018, p. 7). The complex and contested nature of the concept of equality manifests in a plethora of diverse, often conflicting, operationalizations of fairness for AI systems, as we will see later in Section 2.5.3.

A relevant interpretation of egalitarianism is John Rawls's principle of *fair equality of opportunity*, which in essence states that "those with similar abilities and skills should have similar life chances" (Rawls, 1999, p. 63). This implies, for instance, that "if Smith and Jones have the same native talent, and Smith is born of wealthy, educated parents of a socially favored ethnicity and Jones is born of poor, uneducated parents of a socially disfavored ethnicity, then if they develop the same ambition to become scientists or Wall Street lawyers, they will have the same prospects of becoming scientists or Wall Street lawyers" (Arneson, 2015, p. 1). From this, we can deduce

that egalitarianism does not invariably preclude differential treatment. In fact, as we have seen earlier, justice also asks for proportional treatment of individuals when they are *not* alike, which implies an *unequal* distribution of whatever is at issue (D. Miller, 2021).

A fundamental question that arises is then: when are inequalities in accordance with justice? Rawls (1999), for instance, argues that a departure from equality is only justifiable if inequalities "improve everyone's situation, and especially the situation of the worst-off" (Wenar, 2021, p. 1). Another perspective introduces the concept of *responsibility-sensitive egalitarianism*, which accounts for the fact that certain individuals "may have acted in ways that appear to qualify them to receive more (or less) of whatever benefit is being distributed" (D. Miller, 2021, p. 1). Relatedly, *luck egalitarianism* (Anderson, 1999) allows such inequalities "resulting from people's free choices and informed risk-taking, but disregard those which are the result of brute luck" (Binns, 2018, p. 7). Conversely, this means that any inequalities resulting from factors that are beyond an individual's control (e.g., being born with a disability, or being a first-generation college student) should be compensated for. D. Miller (2021, p. 1) notes that one of the main weaknesses of luck egalitarianism is the difficulty of *quantifying* the disadvantage due to brute luck in ways that a "compensatory scheme could be scheduled."

Implementing such a compensatory scheme is also intimately connected with the policy of *affirmative action* (Holzer & Neumark, 2000; Noon, 2010), which involves the deliberate recognition of certain characteristics that are considered to have disadvantaged a group of people through no direct fault of their own—for instance, gender, disability status, race/ethnicity, sexual orientation, or age. In the words of Noon (2010, p. 730), "it brings consideration of the disadvantage into the formal decision-making process by making these characteristics legitimate criteria for evaluating candidates." However, the legitimacy of affirmative action has been subject to fierce debates (Rubenfeld, 1997), and in fact, the US Supreme Court has recently ruled that colleges and universities in the US can no longer consider applicants' race in their admissions procedures (de Vogue et al., 2023). In Section 2.5.4, we will see that this idea of scrapping sensitive information has an equivalent in the FAccT literature, albeit with severe limitations.

2.5.2 The AI Systems Perspective

In the wake of recent controversies such as the 2016 COMPAS debate (Angwin et al., 2016), discussions surrounding justice and fairness have broadened to encompass

the field of AI (Hutchinson & Mitchell, 2019). This is logical, considering that the AI systems of interest in this thesis do not operate in a vacuum. They are sociotechnical systems, devised and deployed by humans, relying on data generated and gathered by humans, and ultimately impacting human lives (Barocas et al., 2019; Ehsan & Riedl, 2020). A common perception (still) holds that AI systems are impartial entities capable of making unbiased decisions (C. C. Miller, 2015). As such, they are often deemed an "evidence-based alternative to biased and idiosyncratic human decisions" (De-Arteaga et al., 2022, p. 3752). However, this is not generally true—unless we assume that the possession of a certain mental state (e.g., systematic animosity against some demographic groups), which we cannot attribute to AI systems, is a prerequisite for wrongful discrimination (Binns, 2018).

AI systems can inflict harm in various ways, stemming from different kinds of algorithmic biases (Jernite, 2022; Olteanu et al., 2019). For instance, Obermeyer et al. (2019) found that a state-of-the-art AI system for predicting healthcare needs was systematically underestimating the sickness of Black patients, thereby depriving them of necessary assistance. This bias arose because the system was predicting healthcare costs as a proxy for sickness, not taking into account that less money is spent on healthcare for Black patients versus White patients due to differences in access to care. Notably, AI systems can not only mirror societal biases ingrained in their training data, but they can *amplify* these biases and harmful tendencies when deployed at scale (Hooker et al., 2020; Jernite, 2022). We will follow up on this observation shortly.

Fairness in AI-informed decision-making typically refers to model outcomes "systematically deviating from statistical, moral, or regulatory standards" (De-Arteaga et al., 2022, p. 3749). It is particularly problematic when such outcomes result in decisions that are biased against people based on their membership in certain groups of the population (Danks & London, 2017; Mitchell et al., 2021). This can result in these groups being unfairly denied desirable opportunities (Barocas et al., 2019), including loans, insurance, or employment (Binns, 2018). When reasoning about discrimination based on group membership, we usually refer to groups that have been subject to "unjustified and systematically adverse treatment in the past" (Barocas et al., 2019, p. 52). The corresponding social categories (e.g., race or gender) are hence commonly referred to as *protected* or *sensitive*. These categories are not just morally sensitive but in many cases also protected by law. The General Equal Treatment Act (GETA) in Germany, for instance, aims to "prevent or [...] stop discrimination on the grounds of race or ethnic origin, gender, religion or belief, disability, age or sexual orientation" (Federal Anti-Discrimination Agency, 2006, Section 1). While the GETA was not initially designed for AI-informed decision-making,

accountability of AI systems at the European level is partially directed by the General Data Protection Regulation (GDPR) as well as the upcoming AI Act (Madiega, 2021). Similar anti-discrimination and consumer protection laws exist in other countries as well (Barocas & Selbst, 2016; Zuiderveen Borgesius, 2018).

Unfairness in AI-informed decision-making may manifest in different ways. Jernite (2022) list four different aspects in which AI systems can inflict harms on humans. Concretely, AI systems can (i) lock in biases and hinder social progress; (ii) spread harmful tendencies even beyond the context of the original training data; (iii) exacerbate existing inequities by over-focusing on stereotypes; and (iv) deny possibilities for recourse by obfuscating biases. All these harms are real. With respect to point (i), Bender et al. (2021) present an argument that large language models might solidify certain values that do not take into account evolving norms, language usage, and communication methods. This can potentially perpetuate older and less inclusive perspectives. Regarding (ii), even if training data is representative of real-world distributions (i.e., there is no sampling bias), AI systems may harm minority groups when optimizing for aggregate performance (De-Arteaga et al., 2022). This has been shown to be problematic in skin cancer detection, where AI systems may only learn predictive symptoms for people with light skin color (Adamson & Smith, 2018). AI systems have also been shown to not only perpetuate but *exacerbate* biases (iii). Hooker et al. (2020), for instance, show how pruning techniques for deep neural networks can have such effects. In a similar vein, Kleinberg and Mullainathan (2019) show that simple models invariably transform historical disadvantages into bias against disadvantaged groups. Notably, they posit that these detrimental effects could be exclusive to simplicity interventions and might not be present in more complex classifiers. Lastly, in terms of point (iv), Durán and Jongsma (2021) propose that AI systems could potentially undermine patient autonomy in clinical decision-making. For instance, these systems might encourage a trend towards paternalistic healthcare practices (McDougall, 2019) by determining the principles under which decisions are made (e.g., prioritizing lifespan extension), which might not align with the patients' values.

All these harms have distinct causes, resulting from biased outputs of AI systems or other parts of the relevant sociotechnical environment (Dolata et al., 2022). Hence, before we can discuss how to detect and, perhaps more importantly, mitigate unfairness, we must understand how potentially harmful biases can infiltrate AI systems. In the following, we use the taxonomy of De-Arteaga et al. (2022) and draw from related works by Danks and London (2017) and Suresh and Guttag (2021).

Bias From Data Collection and Representation This is perhaps the most frequently discussed source of bias. From a fairness perspective, it is vital to consider quality over quantity of data, ensuring it accurately represents the population one aims to make decisions for. This relates to *who* is included in training data. Sampling bias occurs when the data does not adequately represent the intended population, which can significantly impact the fairness of AI systems trained on such data. This is especially true for marginalized groups who have historically been excluded or underserved and are often underrepresented in data. For instance, clinical trials for personalized treatment plans predominantly involve White patients (Warren et al., 2020). Thus, the derived treatment rules may not be effective when applied to a diverse population including patients of other races. Additional examples can be found in the fields of facial recognition (Buolamwini & Gebru, 2018) and widely used datasets for computer vision tasks (K. Yang et al., 2020). Crucially, as De-Arteaga et al. (2022) point out, having representative training data does not suffice to prevent harmful downstream effects, particularly when AI systems are optimized for predictive accuracy, and when accuracy on minority populations does not significantly affect the overall performance.

Not only the selection of observations but also the choice of features can be crucial for fairness considerations, especially when certain features exhibit differing predictive power across demographic groups (Corbett-Davies & Goel, 2018). This corresponds to *how* observations are represented in training data. Often, system developers might choose to measure and utilize information in predictive tasks that are indicative of an outcome for specific groups, while the same features may have minimal predictive or prescriptive power for minorities. This can ultimately result in an inferior performance for minority groups. Such occurrences have been noted in diverse sectors, including healthcare, where certain symptoms might be predictive of skin cancer for White individuals but not for Black people (Adamson & Smith, 2018). Similarly, in the domain of college admissions, considering results from standardized admission tests might inadvertently favor candidates from wealthier backgrounds who can afford additional educational resources such as tutoring lessons (L. Hu et al., 2019).

Lastly, the labels used for training AI systems may be biased—referring to *what* is being predicted. The first issue that may arise is a disconnect between what one aims to measure and what is actually predicted. In such instances, the labels are merely proxies for the outcome of interest, and these proxies might be of varying quality (Bastani, 2021). A key example of this bias is highlighted by Obermeyer et al. (2019), who found that AI systems predicting healthcare needs unfairly disadvantage Black individuals. The issue arises because the predicted proxy outcome is healthcare

expenditure, which tends to underestimate healthcare needs for Black people due to their lower costs stemming from limited access to healthcare resources in the first place. Moreover, labels may be biased if they are generated by humans. For instance, if an AI system learns from historical data reflecting job acceptance or rejection decisions, the system might pick up and perpetuate these biases (Dastin, 2018).

Bias From Model Estimation Biases may infiltrate AI systems even after the training data has been fixed. Favoring overall accuracy, a common performance metric in AI-informed decision-making, can inadvertently neglect minority groups, even when there is no sampling bias. In predictive modeling, the quest to prevent overfitting often involves the use of constrained classifiers or the incorporation of loss functions that impose penalties on complexity (Hooker et al., 2020; Kleinberg & Mullainathan, 2019). If different sub-populations are unequally represented, observations from the majority group could disproportionately influence the final functional representation of a classifier, potentially resulting in models that do not generalize effectively to minority populations (De-Arteaga et al., 2022). This suggests that despite proportional representation of different subgroups in the data, the drive for overall performance optimization can still result in algorithmic bias. On a broader level, it is crucial to note that some tasks are inherently biased and unethical, and thus, they should not be pursued. Examples of such ethically questionable tasks include predicting gender from facial features, which could disproportionately impact vulnerable communities such as transgender individuals (Hamidi et al., 2018); inferring emotions, which may violate privacy (Prégent, 2022); or making assumptions about personality traits and sexual orientation based on physical appearance (Cai & Liu, 2022; McFarland, 2016; Y. Wang & Kosinski, 2018).

Bias From Deployment Finally, it is important to note that even if an AI system's output is unbiased, these systems often do not operate with complete autonomy. As noted in Section 2.2, there is a prevalent argument among researchers and policymakers that humans should maintain ultimate decision-making authority. In such circumstances, AI systems offer guidance rather than definitive decisions, and their output becomes one factor among others that a human decision maker evaluates. When AI systems are used in this manner, the fairness of the ultimate decisions largely hinges on how humans interact with and rely on these recommendations. While the extremes of algorithm aversion and automation bias (i.e., cases where humans tend to reject or accept too many AI recommendations) are well-documented, our understanding of the mechanisms through which reliance behavior impacts fairness in human-in-the-loop decision-making is still underdeveloped. This gap

exists despite the clear evidence of disparities in reliance behavior across different demographic groups, which may result in unfair outcomes, as observed in criminal justice scenarios (Skeem et al., 2020), among others. Studying the interplay of reliance behavior and distributive fairness will be a prominent focus of Part III of this thesis. In the following section, we cover different approaches to measure and mitigate harmful biases in AI systems, as they have been proposed in the pertinent literature.

2.5.3 Statistical Fairness Notions

First, we need to establish some notation. Revisiting the technical preliminaries outlined in Section 2.1, we define f as a classifier, which processes a set of features X and generates a prediction $f(X) = \widehat{Y}$, aiming to estimate the true outcome Y. In most cases, we are concerned with decisions that involve either a desirable (✓) or a non-desirable (✗) outcome. Note that this assumption is reasonable in many scenarios but may not hold in others, such as recommender systems, where preferences between individuals may be vastly different (Binns, 2018). In lending, for instance, the desirable outcome typically corresponds to being granted a loan, whereas the non-desirable outcome is to be denied. Hence, we mostly consider binary classification setups, where $Y, \widehat{Y} \in \{✓, ✗\}$. Moreover, we introduce A as a sensitive attribute, which—for clarity of exposition—takes on two values, a or b, where a indicates membership of the disadvantaged group, and b indicates membership of the advantaged group. We may think of A as referring to gender or race. Having established these preliminaries, we now discuss how FAccT researchers have been thinking about detecting and mitigating unfairness in AI-informed decision-making.

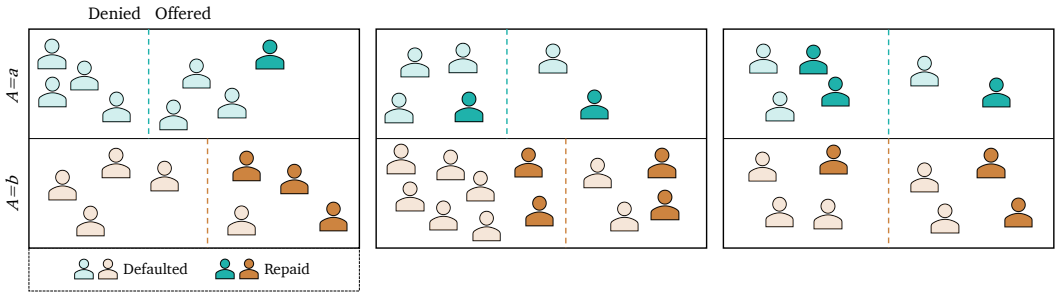
Detecting Unfairness The FAccT community has proposed a variety of statistical fairness notions that can be used to measure (un)fairness in AI-informed decision-making. Most of them map to some broader idea of (mostly distributive) justice and are informed by current laws and regulations (Barocas & Selbst, 2016). Generally, statistical fairness notions aim to define non-discrimination in the form of statistical expressions incorporating random variables that describe a classification or decision-making situation (Barocas et al., 2019). Formally, these criteria are characteristics of the joint distribution of A, Y, $\widehat{Y}$, and occasionally the features X. This implies that we can conclusively determine whether a fairness criterion is fulfilled by examining the joint distribution of these random variables (Barocas et al., 2019). It also means that in many cases we can quantify deviations from fairness through statistical

metrics, allowing us to reason about the *degree* of fairness or unfairness of a given system. Before we proceed, we must underscore that fairness is a social and ethical concept (Chouldechova, 2017), which is complex and contested (Mulligan et al., 2019). As such, it can never be universally quantified or confined within a single metric. As De-Arteaga et al. (2022, p. 3758) stress, any use of a given fairness metric ought to be carefully weighed and "grounded on context-dependent goals and values that it aims to advance." On this note of caution, we now turn to discussing three statistical notions of fairness that have been prevalent in research and practice.

The first notion is *demographic parity*, which is one of the most widely considered fairness notions in the literature (Mehrabi et al., 2021). Intuitively, demographic parity suggests that positive outcomes (✓) should be equally probable across different groups. For instance, an AI system used in recruitment adheres to demographic parity if it offers job opportunities equally to both men and women. That is, if 50% women and 50% men apply at a company, then the pool of candidates who receive a job offer should consist of 50% women and 50% men. This is also illustrated in Figure 2.4 (a) on page 47. Statistically, demographic parity requires that $\widehat{Y}$ is independent of the sensitive attribute A, or, equivalently, that $P\left(\widehat{Y} = ✓ | A = a\right) = P\left(\widehat{Y} = ✓ | A = b\right)$ holds. It is often argued that demographic parity embodies a presumption of equality, implying that all groups should have an equal right to acceptance (Hardt et al., 2016). However, note that demographic parity does not consider the true outcome Y and, hence, cannot account for any possible correlations between Y and the sensitive attribute A. While this may not pose any problems in some cases, it can be undesirable in others (De-Arteaga et al., 2022).

The second notion of *equalized odds* requires that an AI system makes type I (false-positive) and type II (false-negative) errors at equal rates across different groups. It is therefore also referred to as *error rate parity* (Hardt et al., 2016). Formally, this means that both $P\left(\widehat{Y} = ✓ | A = a, Y = ✓\right) = P\left(\widehat{Y} = ✓ | A = b, Y = ✓\right)$ and $P\left(\widehat{Y} = ✓ | A = a, Y = ✗\right) = P\left(\widehat{Y} = ✓ | A = b, Y = ✗\right)$ must be met. An example of equalized odds is shown in Figure 2.4 (b) on page 47. One might wonder what the normative justification for equalizing error rates—that is, for satisfying equalized odds—of an AI system is. A possible answer is that equalized odds brings attention to the question of who should bear the costs of misclassification (e.g., to be wrongfully denied a loan). In that respect, equalized odds may be thought of as a way to equalize the burden of misclassification on different groups (Barocas et al., 2019). By explicitly considering Y, equalized odds also accounts for a "sense of merit" (Barocas et al., 2019, p. 57), which in the context of lending refers to some measure of creditworthiness. Depending on the context and the associated costs of different errors, it is sometimes meaningful to only demand that *either* type

Fig. 2.4.: Illustration of different statistical fairness notions.



(a) Demographic parity (b) Equalized odds (c) Predictive parity

Note: Inspired by De-Arteaga et al. (2022), we illustrate three statistical notions of fairness in AI-informed decision-making for the case of financial lending. Horizontal lines separate the disadvantaged (top) from the advantaged group (bottom). Vertical dashed lines indicate the decision threshold for each group: applicants to the left are denied, applicants to the right are offered a loan. The first notion of demographic parity *(a) requires that the percentage of applicants that are offered a loan be equal across groups. This is the case in our example because the bank offers a loan to 50% of applicants in both the advantaged and the disadvantaged group. The notion of* equalized odds *(b) requires that equal shares of creditworthy applicants in both groups be offered the loan (here: 50%), and also that equal shares of uncreditworthy applicants be offered the loan (here: 25%). Finally, predictive parity is satisfied in example (c) because the default rate (i.e., applicants that are granted the loan but cannot pay it back) is equal across groups at 50%.*

I *or* type II errors be equalized (Barocas et al., 2019). While it is easy to argue that being wrongfully denied a loan involves a cost for the respective individual, being wrongfully *granted* a loan, however, may not be a burden at all—at least in the short run.

The third statistical fairness notion that is often applied is *predictive parity*, and it is closely related to the idea of *calibration* (Barocas et al., 2019). It demands that the probability of a correct prediction is equal across values of a sensitive attribute (Chouldechova, 2017). Figure 2.4 (c) illustrates a case where predictive parity is satisfied: whenever the AI system recommends a loan, probability of default is the same across both groups ($\frac{1}{2} = \frac{2}{4} = 50\%$). Predictive parity is another conditional independence statement, requiring that $P\left(Y = \checkmark | A = a, \widehat{Y} = \checkmark\right) = P\left(Y = \checkmark | A = b, \widehat{Y} = \checkmark\right)$ and $P\left(Y = \checkmark | A = a, \widehat{Y} = \text{✗}\right) = P\left(Y = \checkmark | A = b, \widehat{Y} = \text{✗}\right)$ hold. De-Arteaga et al. (2022) note that predictive parity might be desirable in risk assessment scenarios, where it ensures that calculated risk retains a consistent meaning across demographic groups. In conclusion, we summarize the three predominant statistical notions of fairness as follows:

Common statistical fairness notions are:

- *Demographic parity*: $P\left(\widehat{Y} = ✓ | A = a\right) \stackrel{!}{=} P\left(\widehat{Y} = ✓ | A = b\right)$
- *Equalized odds*: $P\left(\widehat{Y} = ✓ | A = a, Y = y\right) \stackrel{!}{=} P\left(\widehat{Y} = ✓ | A = b, Y = y\right),\ y \in \{✓, ✗\}$
- *Predictive parity*: $P\left(Y = ✓ | A = a, \widehat{Y} = y\right) \stackrel{!}{=} P\left(Y = ✓ | A = b, \widehat{Y} = y\right),\ y \in \{✓, ✗\}$

The three fairness notions we have discussed all revolve around the concept of ensuring fairness among groups, which is sometimes referred to as *group fairness* (Binns, 2020). Other perspectives emphasize *individual fairness*, stipulating that individuals who are similar in a specific context ought to be treated in the same manner. This notion is commonly known as *fairness through awareness* (Dwork et al., 2012). Typically, the similarity between individuals is evaluated without regard to their group affiliation (De-Arteaga et al., 2022). In real-world scenarios, however, applying individual fairness notions relies on a suitable similarity metric within the feature space, meaning that one has to define what constitutes *similarity* among individuals. Formulating and justifying this metric can often be challenging (De-Arteaga et al., 2022). Another set of methodologies striving to identify unfairness relies on causal inference. Kusner et al. (2017) introduce the concept of *counterfactual fairness*, which postulates that any causal impacts of a sensitive attribute on an outcome are ethically unjustifiable. These causal approaches typically rest on strong assumptions related to counterfactual estimation. L. Hu and Kohler-Hausmann (2020) even argue that their underlying ontological assumptions are fundamentally flawed, given that many of the effects attributed to sensitive attributes (e.g., gender or race) are actually intrinsic characteristics of these attributes as a social status.

Given the multitude of ideas on how to operationalize fairness in AI-informed decision-making, it may naturally lead one to question if it is possible to incorporate *all* these notions. Unfortunately, it has been shown that this is generally not feasible, as different notions of fairness often conflict with one another (Chouldechova, 2017; Kleinberg et al., 2017). At the same time, empirical studies by Bell, Bynum, et al. (2023) and others suggest that it might be possible to reconcile some notions, for instance, by permitting minor degrees of relaxation. Notably, Binns (2020) argues that some conflicts (e.g., between group and individual fairness) are solely artifacts of the technical operationalization of different notions of fairness, and that these conflicts are not present in the normative underpinnings of justice in moral and political philosophy. Regardless, different fairness notions mirror diverse worldviews and interpretations of justice, and the selection of a suitable metric should be

considered based on various factors and specific to the context at hand. To assist in choosing an appropriate metric, De-Arteaga et al. (2022) provide some guidance by weighing the advantages and disadvantages of different fairness notions, including considerations of how the choice of a particular metric might impact other desiderata in a given decision-making scenario.

Mitigating Unfairness We have previously seen that harmful biases can infiltrate AI systems in many different ways. We also discussed that the definition of *fairness* is subjective and may depend on individual worldviews, cultural backgrounds, and societal norms. This, however, does not imply that mitigating unfairness in AI-informed decision-making is a lost cause. We now discuss different techniques to address problems of unfairness. These methods are typically categorized based on the stage of the AI system lifecycle at which we intervene. Therefore, the FAccT literature distinguishes three different categories of mitigation techniques: *pre-processing*, *in-processing*, and *post-processing*.

Pre-processing typically refers to methods used to modify existing datasets, but it may also include strategies for gathering high-quality data from the outset. The earliest point of intervention is the collection of superior training data, which helps prevent the introduction of sampling and representation biases, as discussed in Section 2.5.2. If collecting new data is not an option (e.g., due to the large amounts of data required for training AI systems), it is of paramount importance to understand how available data was generated and to identify any potential biases. Adherence to reporting standards, like those suggested by Gebru et al. (2021), can enhance this understanding. If there are concerns about certain aspects of the training data, many techniques exist to manipulate that data in a way that a classifier may not pick up harmful biases at this stage. Common methods involve various forms of resampling, such as augmenting the proportion of observations from underrepresented groups (Kamiran & Calders, 2012). Zemel et al. (2013) also suggested learning "fair representations" of the data that conceals information about individuals' affiliation with protected groups. However, as noted in Section 2.5.2, harmful biases can infiltrate AI systems at later stages, even if training data appears to not suffer from any problematic biases.

In-processing techniques aim to incorporate fairness considerations during the AI system development. If a given system is based on ML, this equates to the model training phase (Mehrabi et al., 2021). In those cases, the idea is to modify the learning algorithm in a way that it not only minimizes the prediction error but also

reduces unfairness in predictions. This could be achieved by adding fairness constraints (e.g., demographic parity) into the optimization problem that the learning algorithm is solving (Zafar et al., 2019), or by adding the fairness constraint as a regularization term directly in the objective function of the problem (Kamishima et al., 2012). Other in-processing approaches involve adversarial learning techniques (B. H. Zhang et al., 2018), where the idea is to maximize a classifier's ability to predict the true outcome Y, while simultaneously minimizing an adversary's ability to predict a given protected attribute A. All in-processing approaches require, by definition, modifications to standard learning algorithms and may not be applicable to all types of models or fairness definitions.

When there is no control over the training data or the training process itself, a third category of mitigation techniques comes into play, known as *post-processing* methods. These methods often involve adjusting classification thresholds for various demographic groups (Kamiran et al., 2012). Some techniques focus on post-hoc calibration of a classifier to align with specific fairness notions (Pleiss et al., 2017), or they might solve an auxiliary optimization problem based on a given model's output, aiming for solutions that are both accurate and fair (Hardt et al., 2016). A significant advantage of most post-processing approaches is their versatility—they can be applied to any classifier, regardless of access to its internal workings or the data it was trained on (Barocas et al., 2019). This flexibility makes them particularly practical in many real-world applications.

2.5.4 The (Flawed) Idea of "Fairness Through Unawareness"

It may seem intuitive to pursue fairness by simply scrapping all sensitive features. This idea is sometimes referred to as "fairness through unawareness" (Kusner et al., 2017, p. 2), and it deems an AI system fair if it does not *actively* consider any sensitive information in the decision-making process. For instance, to prevent racial bias, one might wish to deny the model access to variables representing individuals' race. Grgić-Hlača, Zafar, et al. (2018) even associate this strategy with the notion of procedural fairness. According to Nyarko et al. (2021), this may be justifiable under a *deontological* account, which judges the fairness of a process irrespective of its consequences. Such a justification might consider it "fundamentally unethical [...] to condition the allocation of costs and benefits on an individual's [gender or race]" (Nyarko et al., 2021, p. 2). However, the inclusion of sensitive attributes does not inherently lead to distributive unfairness (Kleinberg et al., 2018), and when it does, exclusion alone might not resolve the issue (Dwork et al., 2012). In some instances, it may even be harmful to omit sensitive information: Barocas

et al. (2019), referring to Bonham et al. (2016), note that certain medication may depend on race in legitimate ways. Therefore, enforcing a disconnect between the prescription of such medications and race could potentially harm individuals. We elaborate more on the relationship between the use or disuse of sensitive features and distributive fairness in the following.

On the one hand, access to sensitive features can enhance an AI system's fairness properties when the predictive relationship between features and outcome differs among subgroups. The following examples are taken from De-Arteaga et al. (2022). Consider, for instance, AI-informed hiring based on educational variables. Socioeconomic disparities and access barriers may result in first-generation college graduates, who possess equal potential for success, graduating with a lower average grade. An AI system that acknowledges whether an individual is a first-generation graduate can account for this bias by recognizing that the predictive relationship between grades and future job performance depends on this socioeconomic factor.

On the other hand, an AI system's access to sensitive features may result in these being used as proxies for the target outcome, which can exacerbate existing disparities. For instance, if structural challenges have affected first-generation students' ability to graduate college or thrive in their first job, an AI system may learn a negative association between being a first-generation graduate and job performance. Consequently, it might predict that a first-generation graduate is less likely to succeed, even if their profile is otherwise identical to others. In such cases, simply excluding the sensitive attribute(s) is often insufficient. Many variables can act as proxies as they correlate with sensitive attributes. This is also referred to as *redundant encoding* (Dwork et al., 2012), which means that sensitive information may be encoded in other (seemingly innocuous) features. For example, a student's zip code or school district may be used by an AI system as a proxy for race. A notable example where excluding sensitive attributes was ineffective is Amazon's rollout of free same-day delivery service in the US. Despite Amazon's claim that no race data was used to decide where to offer free delivery services, other variables (such as the zip code) acted as potential proxies, leading to a service that primarily excluded Black neighborhoods (Ingold & Soper, 2016).

2.5.5 Fairness Perceptions

It is often posited that mitigating unfairness in AI-informed decision-making cannot solely rely on technical solutions, but also necessitates methodologies rooted in the social sciences and humanities (Barabas et al., 2020; Sloane & Moss, 2019).

More specifically, Starke et al. (2022, p. 2) insist that gaining a "thorough empirical understanding of when and why citizens perceive [AI systems] to be (un)fair" is vital in the creation of human-centered AI systems. This perspective has also been dubbed as "society-in-the-loop" by Rahwan (2018). Accounting for fairness perceptions is particularly important in human-in-the-loop decision-making: while technical interventions aim at mitigating harmful biases in the AI system, they do not account for the human decision maker, who has discretionary power to override the system. The way these decision makers perceive the fairness of an AI system may influence their reliance on AI recommendations, which can have downstream effects on various metrics of decision quality and distributive fairness. We study this relationship between perceptions and distributive fairness in Chapter 8 of this thesis. Moreover, understanding human perceptions is vital even when an AI system is considered "fair" based on certain statistical fairness notions. This is because a given statistical notion of fairness may not necessarily align with how humans, especially decision subjects, perceive fairness (Saxena et al., 2020).

An important question arises: how can we measure these fairness perceptions? Similar to statistical notions of fairness, there is no one-size-fits-all metric to quantify perceptions of fairness (Starke et al., 2022). A significant portion of the relevant research has utilized existing constructs, primarily from the field of organizational justice (Colquitt et al., 2001; Greenberg, 1987). These constructs, originally intended for use in workplace contexts, such as the relationship between an employee and their manager, have been adopted by researchers in AI and adjacent fields to gauge human perceptions towards AI systems (Starke et al., 2022). The literature distinguishes different facets of fairness perceptions, mirroring our discussion on various dimensions of justice in Section 2.5.1. When applied to AI systems, *distributive fairness* refers to the outcomes generated by AI systems, while *procedural fairness* aspects concern the decision-making procedures an AI system employs to reach an outcome. Most often, these constructs of fairness perceptions are measured on Likert scales, either through single or multiple items.

The organizational justice literature also identifies other dimensions, including *informational fairness*, which have not been extensively studied in the realm of AI systems. Binns et al. (2018), one of few exceptions, measure informational fairness by asking humans whether they understand the process by which an AI-informed decision was made. Intuitively, informational fairness refers to how adequately an AI system explains its actions (D. Chan, 2011). Colquitt and Rodell (2015) propose several measurement items for informational fairness in the workplace context, including how candid, thorough, reasonable, timely, and personalized the communication and explanation of a decision is. Despite often being overlooked in

AI research, we emphasize the importance of the informational fairness construct as a link between the two main themes of this thesis: transparency and fairness. We will focus on this in more depth in Chapters 5 and 6.

While *trustworthiness* is not technically a component of fairness perceptions, we cover it as well, because it is commonly measured in experimental studies on the effects of transparency mechanisms. Trust is a somewhat nebulous concept, with varying definitions and measurements that are often inconsistent (Jacovi et al., 2021; Papenmeier et al., 2022). Some define *trust* as the degree to which the trustee believes that an AI system will behave as anticipated (Gol Mohammadi et al., 2013; Papenmeier et al., 2022). Others stress the importance of vulnerability of the trustee (Mayer et al., 1995) and the expectation that the AI system will abide by an implicit or explicit agreement (Jacovi et al., 2021). Such terminological inconsistencies make it difficult to compare empirical findings. It is also important to distinguish between *trust* and *reliance*, as they are often mistakenly used interchangeably. In the context of human-in-the-loop decision-making, *reliance* refers to the behavior of adhering to or overriding AI recommendations (Lai et al., 2021; Schemmer et al., 2023). *Trust*, on the other hand, is a subjective attitude towards the entire AI system, which evolves over time (Parasuraman & Riley, 1997; Rempel et al., 1985; K. Yu et al., 2017). Common dimensions of fairness perceptions are summarized in the following. We stress that these constructs are nuanced and may not be measured through single items. Hence, this overview is not meant to serve as a universal definition but merely a loose guide for explanatory purposes.

Common dimensions of fairness perceptions are:

- *Distributive fairness*: Outcomes of an AI system are fair
- *Procedural fairness*: Decision-making processes of an AI system are fair
- *Informational fairness*: Explanations of an AI system are adequate
- *Trustworthiness*: An AI system is reliable and of integrity

Prior work has been trying to understand both algorithmic and human predictors of fairness perceptions (Starke et al., 2022; Toussaint et al., 2022). Saxena et al. (2020) assess how humans perceive different statistical notions of fairness in the context of lending, suggesting that study participants preferred a notion of meritocratic fairness (Joseph et al., 2016), according to which individuals receive an amount of money that is proportional to their respective repayment rates. M. K. Lee et al.

(2019) empirically show that fairness perceptions decline for some people when their personal fairness concepts differ from those of the AI system. These findings are of relevance to this thesis because they hint at potential discrepancies between statistical notions of fairness and what humans perceive as fair. A large fraction of studies have also examined the effects on fairness perceptions in the presence of different transparency mechanisms. These empirical findings are mostly inconclusive, stressing that fairness perceptions depend on many factors, such as the explanation style (Binns et al., 2018; Dodge et al., 2019), the use case (Angerschmid et al., 2022), user profiles (Dodge et al., 2019), or the decision outcome (Shulner-Tal et al., 2022; R. Wang et al., 2020). Some work has also assessed the impact of people's demographics, including gender (Pierson, 2017), as well as political views and task experience (Grgić-Hlača et al., 2022) on their perceptions.

A series of relevant prior studies have found that knowledge about the features that an AI system uses influences people's fairness perceptions. Specifically, people tend to be averse to the use of sensitive information, such as gender or race (Corbett-Davies & Goel, 2018; Grgić-Hlača, Redmiles, et al., 2018; Grgić-Hlača, Zafar, et al., 2018; Grgić-Hlača et al., 2020; Nyarko et al., 2021; Plane et al., 2017). Nyarko et al. (2021) conduct several empirical studies and also observe that people are generally averse to the use of gender and race in AI-informed decision-making. Interestingly, people's perceptions towards these features change after they learn that "blinding" the AI system to these features can lead to *worse* outcomes for marginalized groups. Similarly, it has been shown that people's perceptions towards the inclusion of sensitive features switch when they are told that this inclusion makes an AI system more accurate (Grgić-Hlača et al., 2020) or equalizes error rates across demographic groups (Harrison et al., 2020). These findings suggest that fairness perceptions are brittle and may sometimes be based on wrong assumptions. Surprisingly, few works have examined downstream effects of fairness perceptions on AI-informed decisions, such as reliance behavior. This will be a major theme of Chapter 8. For an in-depth review of empirical findings on fairness perceptions, we refer to Starke et al. (2022).

3 Fairness Desiderata of Transparency

In this chapter, we conduct a systematic review of pertinent literature exploring the complex interplay between transparency and fairness in artificial intelligence (AI)-informed decision-making. From this comprehensive review, we distill nine canonical claims frequently made concerning the relationship between transparency and fairness. A significant proportion of these claims view transparency as a pathway to fairness, while a few express concerns about transparency potentially undermining fairness. Notably, we discover that many of these claims are not substantiated by empirical evidence, casting doubt on the role of existing transparency mechanisms as an ethical panacea.

3.1 Introduction

With the aim of enhancing accuracy, efficiency, and objectivity, AI systems are increasingly being utilized in a wide range of recommendation and decision-making tasks, permeating various aspects of our daily lives. These systems now play a role in shaping healthcare practices (K.-H. Yu et al., 2018), influencing career choices (Upadhyay & Khandelwal, 2018), predicting credit scores (N. Chen et al., 2016), and assessing criminal risk (Angwin et al., 2016). While AI surpasses many human limitations, it also introduces new risks, particularly when deployed in socially sensitive or high-stakes contexts. One such risk is the propagation of historical biases, leading to potential discrimination against specific demographic groups (Barocas & Selbst, 2016). Additionally, the remarkable progress achieved

This chapter is based on in-progress work as follows:

> Schöffer, J., Deck, L., De-Arteaga, M. & Kühl, N. (2023). Overcoming intuitions: A critical survey on fairness benefits of explanations. *Working Paper*.

in AI often comes at the expense of complexity and opacity, hindering human comprehension (Burrell, 2016), especially for non-experts (Laato et al., 2022).

Researchers and policymakers have expressed concerns about opacity and societal consequences of AI systems (Cath, 2018; Floridi et al., 2018; Selbst & Barocas, 2018). A prevalent assumption is that transparency and fairness are intrinsically linked, with transparency functioning as a facilitator or even a prerequisite for achieving fair AI systems (Arrieta et al., 2020; Gill et al., 2020). However, despite the abundance of claims supporting this premise, empirical evidence on this purported relationship remains scarce (Balkir et al., 2022). Furthermore, both transparency and fairness are contested concepts, encompassing a wide range of ideas and being employed in diverse contexts to address various stakeholders (Langer, Oster, et al., 2021).

This study aims to critically examine claims regarding the potential of transparency mechanisms to promote fairness, taking into account the available empirical evidence. We perform a systematic literature review across Scopus and arXiv.org, resulting in a corpus of 169 papers. In each paper, we assess claims pertaining to the relationship between transparency and fairness, and we evaluate their grounding. Through qualitative analyses, we identify nine canonical claims as a foundation for our analysis. Among these claims, seven address assumed capabilities of transparency mechanisms with respect to fairness (we refer to them as *desiderata*), while two engage in fundamental debates about how shortcomings of existing transparency mechanisms might compromise the pursuit of fairness. Our findings reveal that evidence across various domains is ambiguous, and claims portraying transparency as an ethical panacea are misleading. In fact, the diverse range of claims suggests that the relationship between transparency and fairness is complex and multifaceted. We also see that while transparency mechanisms may be helpful for ethically-minded system developers and decision makers in certain cases, evidence regarding their utility for decision subjects and regulators remains inconclusive. We argue that delineating precise goals and specifying the targeted stakeholders would significantly clarify the potential capabilities of transparency mechanisms. Moreover, we advocate for more judicious use of feature importance techniques and underscore the necessity for a new conceptualization of informational fairness.

The primary contributions of this work are: (i) a systematization of recent fairness-oriented transparency research through the mapping of claims and evidence with a stakeholder-centered view, and (ii) a critical examination of capabilities and limitations of existing transparency mechanisms, as suggested in the current literature. The remainder of this chapter is organized as follows: we begin by briefly reviewing related work, after which we outline the methodology employed in our systematic

literature review and qualitative claim analysis. In the main part of this work, we present and discuss the identified canonical claims as they pertain to transparency as a means or a threat to fairness. Finally, we conclude this chapter summarizing our findings and implications.

3.2 Related Work

Despite the extensive body of research developed independently around transparency and fairness in AI systems, a scarcity of work at their intersection has been noted in recent studies (Balkir et al., 2022; Meng et al., 2022). In a more specific context, Balkir et al. (2022) scrutinize the challenges of employing transparency mechanisms for fostering fairer natural language processing (NLP) models, identifying a fundamental discrepancy between the concepts of transparency and fairness: they argue that transparency predominantly concerns procedural understanding, whereas fairness emphasizes the necessity of equitable outcomes. At a broader scale, Langer, Oster, et al. (2021) evaluate desiderata of transparency in over 100 peer-reviewed studies, revealing that only a subset of these studies validate their propositions with empirical evidence. In line with this, following an exhaustive review, Doshi-Velez and Kim (2017) advocate for a careful alignment between claims and methodological foundations. Moreover, Lipton (2018) questions prevalent perspectives on transparency, stressing the importance of explicit definitions and verifiable objectives of transparency mechanisms.

While prior surveys have explicitly considered the potential of transparency for fairness, they often lack systematic methodology and scarcely discuss any limitations and shortcomings. Abdollahi and Nasraoui (2018) survey explainable recommender systems, accentuating their capacity to detect sources of bias. Zhou et al. (2020) offer an overview of transparency mechanisms addressing fairness issues, underlining the significance of contextual factors and interdisciplinary research. Our study adopts a methodical and critical approach akin to Blodgett et al. (2020), who critically analyze the usage of the term *bias* in the context of NLP. Distinguishing between intuitive claims and various forms of evidence, we aim to scrutinize any discrepancies. Moreover, by incorporating criticism of transparency mechanisms, we intend to mirror the current scholarly discourse, with the hope of contributing an essential step towards greater clarity and specificity in this field.

3.3 Methodology

Drawing upon the methodology from Blodgett et al. (2020), we set out to systematically identify and categorize claims to derive higher-order insights. We deemed a claim as pertinent if it explicitly indicated any impact of transparency in AI on fairness, as per our defined parameters, and if it offered a unique argument that was not simply echoing the findings of prior work. These shortlisted claims then underwent a qualitative analysis. Adapting the process proposed by Wolfswinkel et al. (2013), we blended deductive literature research with inductive coding at the level of individual statements, rather than the broader article level.

To prepare a comprehensive set of search criteria, we first established a preliminary understanding of the research domain and tested various combinations of search strings using Google Scholar, as detailed in Section 3.3.1. Leveraging these preliminary insights, we configured our systematic research approach that led to the identification of 169 papers. We explain this approach in detail in Section 3.3.2. Upon identification, our collection of papers was examined for relevant claims. The final step of our analysis resulted in the emergence of 9 canonical claims, 7 of which pertain to transparency as a means for fairness (we will refer to these as *desiderata*), and 2 as a threat to fairness—this is discussed in Section 3.3.3.

3.3.1 Exploratory Literature Review

To gain an understanding of the research domain, test the efficacy of specific keywords, and ascertain relevant publication venues, we commenced our investigation by querying the Google Scholar database. Our search string aimed to reflect various dimensions of both transparency and fairness, while limiting the search results to AI contexts. Beyond terms such as *explainable AI* and its acronym *XAI*, we consulted Arrieta et al. (2020) and incorporated related terms like *understandability, comprehensibility, interpretability, explainability,* and *transparency*. We also included the commonly used keyword *explanation*. Overall, our goal was to capture the multifaceted nature and different definitions of transparency in the literature. In the context of fairness, we discovered that the term *fair* (and its compound derivatives) appeared to dominate the discourse in all relevant papers. While acknowledging that the terms *discrimination, justice, ethics,* and *bias* are sometimes used interchangeably with *fairness*, we excluded these terms to mitigate unnecessary noise.

To direct our search towards the field of AI, we opted for straightforward terms such as *artificial intelligence* and *machine learning*. We found that related terms such as

algorithm and *automated* did not enhance the quality of results and were therefore excluded. After screening approximately 400 individual papers, we finalized the following search string, noting that the use of the asterisk as a wildcard character allowed us to consider both adjective and noun forms of a term:

> (xai **OR** explanation **OR** understandab* **OR** intelligib* **OR** comprehensib* **OR** interpretab* **OR** explainab* **OR** transparen*) **AND** fair* **AND** (ai **OR** "artificial intelligence" **OR** "machine learning")

In addition, by screening titles and abstracts, we were able to discern relevant literature sources. Of the approximately 400 individual papers reviewed, 155 were deemed useful in addressing our research question. It is important to note that this selection was not the set of papers employed for the review but served merely as a directional guide throughout the process. This preliminary analysis allowed us to determine that ACM (31%) and Springer (19%) emerged as the predominant publishers in the field. Furthermore, IEEE (6%), Elsevier (6%), and AAAI (5%) showed significant presence, marking them as promising sources for our investigation. Recognizing the current and rapidly evolving nature of the topic, our review also includes preprints from arXiv.org. This decision was further motivated by the fact that, in our initial review, approximately 12% of the papers were published on this platform. Other sources contributed relatively few papers, with each offering four or fewer relevant publications.

3.3.2 Systematic Literature Review

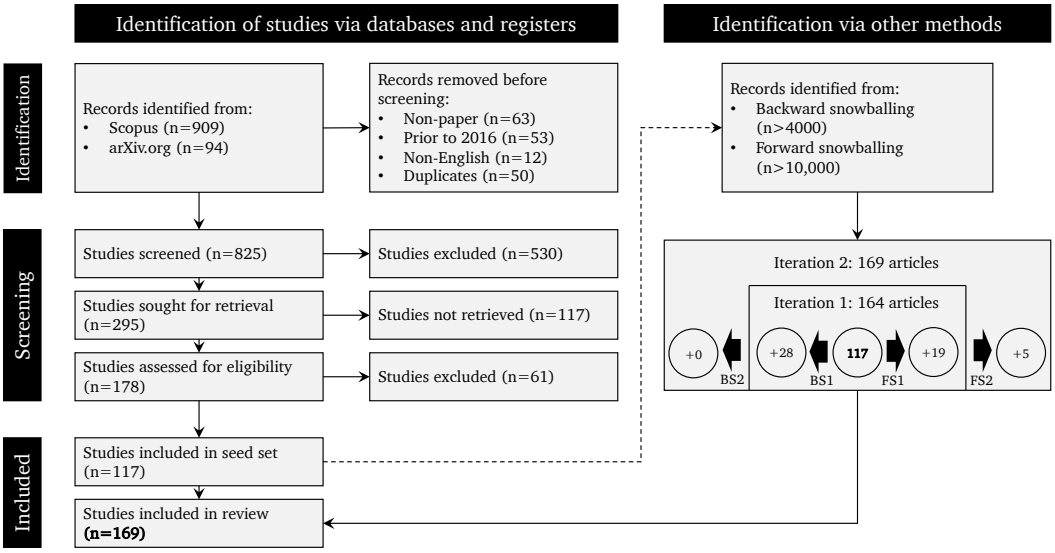
Building upon recent arguments favoring the combination of two prevalent search strategies, database querying and snowballing (Wohlin et al., 2022), we adopted established guidelines for systematic literature reviews in the software engineering domain (Kitchenham & Charters, 2007; Wohlin, 2014). Scopus was our primary choice for database search, given its effectiveness in generating seed sets for snowballing (Mourão et al., 2020), and its inclusion of all relevant publishers identified in Section 3.3.1, except arXiv.org. To include recent, non-peer-reviewed manuscripts, we implemented our search string in the arXiv.org database, albeit limiting the search to keywords due to technical constraints of the search feature.

Our process, following the documentation guidelines of Kitchenham and Charters (2007) and the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) standard (Page et al., 2021), is aimed at maintaining transparency and ensuring replicability. Figure 3.1 on page 61 depicts the multi-stage condensation of a total body of 1003 identified records (as of September 2022) to a core set of 117, detailing the filter criteria applied at each stage. First, we considered only full papers, excluding records like courses, keynotes, and the like. Each abstract was manually inspected to retain only those papers which explored dimensions of both fairness and transparency in alignment with our defined scope. As a result, papers with overly broad or divergent interpretations of transparency (e.g., using the term *explain* in an unrelated context), or those that used *fair* in different contexts (e.g., "fairly") were discarded. We also excluded papers where fairness or transparency were not the primary focus of research.

Upon proceeding to full-text analysis, we heuristically scanned each paper for specific claims pertaining to transparency and fairness. We prioritized unique statements over straightforward summaries or paraphrases of previous work, which largely eliminated literature reviews. Finally, we discarded papers where the direct relationship between transparency and fairness was either not considered or remained excessively vague. For instance, Shin (2020) examines the influence of transparency and fairness on trustworthiness but does not address the interaction of transparency and fairness. Yet, the broad scope and findings of these papers provide valuable insights into the relationship between transparency and fairness, and will be beneficial for subsequent discussions in this thesis as well as prospective considerations.

Initiating with a seed set of 117 papers, we then employed iterative backward and forward snowballing strategies (Wohlin, 2014). The `citationchaser` tool (Haddaway et al., 2022) facilitated the generation of a comprehensive list of all unique references (backward snowballing) and citations (forward snowballing). However, due to `citationchaser`'s limit of 100 starting articles and the absence of 2 entries in the underlying Lens.org database, the remaining articles were manually inspected using Google Scholar. We began by examining the titles, subsequently implementing a filtering procedure similar to the one used in the database search. To identify potentially overlooked relevant literature efficiently, we sorted the results according to their frequency of occurrence in the citation graph, allotting more attention to more frequently referenced or cited papers. Thus, publications appearing only once in the citation graph received minimal consideration. Upon repeating this procedure with the newly added papers from the first iteration, we identified 5 additional pertinent publications. No further relevant results emerged in a third iteration, prompting us to halt the process.

Fig. 3.1.: PRISMA flowchart describing the article selection procedure.



3.3.3 Analysis of Claims

Following the establishment of our literature corpus, we inductively discerned dominant themes by analyzing shared characteristics and clustering claims into meaningful groups. Borrowing elements from grounded theory, we adhered to the research design framework from Chun Tie et al. (2019), utilizing `MAXQDA` for claim extraction, coding, and memoing (Kuckartz & Rädiker, 2019). Our analysis was conducted around the following central question:

> *What claims does recent literature make regarding the relationship between different forms of transparency and fairness in AI systems?*

Our initial step was to skim the complete texts of our selected 169 papers to understand their respective methodologies and main findings. Concurrently, we sought claims with a strong emphasis on the most pertinent sections of the articles. For instance, the introduction, discussion, and conclusion sections often provided more cogent claims than sections regarding methodology. Throughout the coding procedure, we employed memos to note crucial insights, enrich the claims with contextual information (such as textual context, meaning of abbreviations, authors' reasoning, etc.), and document the coders' thought processes. In the first iteration, we preserved the codes as specific as possible to retain a maximum amount of information. While coding, we considered not only the explicit content of the claims but also their

context and, when feasible, their underlying logic. This information was used to categorize the type of evidence that led to the claim, which we documented in our coding system.

Subsequent iterations facilitated the identification of higher-level concepts, and we began to cluster the claims into mutually exclusive categories. To ensure theoretical saturation, we validated the sufficiency of the identified categories and the plausibility and accuracy of the assignment by revisiting each claim. In alignment with the recommendation of Kuckartz and Rädiker (2019) to test for intercoder agreement, we selected a random set of 30 claims and addressed disagreements in category assignment within our team of authors. These discussions affirmed the adequacy of the higher-level codes, resulting in only minor reassignments.

3.4 Findings and Implications

To offer a comprehensive perspective of the set of papers analyzed, we begin by detailing the methodologies employed within these studies. The primary intention hereby is not to achieve a perfectly distinct categorization, but rather to highlight the underlying type of evidence that is fundamental to the insights discussed in Sections 3.4.1 and 3.4.2. In this context, empirical evidence warrants different consideration than conclusions based on intuition or reasoning, and results derived from human subject experiments should be interpreted differently than empirical evaluations on public datasets. Table 3.1 on page 63 breaks down the methodologies utilized in the 169 papers and offers notable examples to elucidate the categories. Note that the counts exceed 169 as some papers employ more than one methodology. For instance, Ahn and Lin (2020) propose a design framework, apply it to real-world data, and conduct user studies to illustrate its usefulness for practitioners.

Conceptual contributions include all studies that do not undertake any primary form of empirical evaluation. This subset encompasses *literature reviews* and *argumentation* (e.g., position papers) which build upon existing work and reasoning. Detailed recommendations for design, evaluation, or regulation, as well as conceptual or formal models also fall into this category, referred to as *frameworks*. The category of *technical* experiments encompasses all studies that empirically evaluate a method or framework on real-world datasets. The most common type of research is the empirical evaluation of a novel transparency or fairness *method*. This also includes studies that scrutinize existing transparency methods by executing adversarial attacks. *Case studies* apply existing methods in a specific domain or context. Moreover,

Tab. 3.1.: Methodologies used in the reviewed papers.

Methodology	Count	Exemplary papers
Conceptual	**73**	
Framework	33	Floridi et al. (2018), Kleinberg and Mullainathan (2019), and Langer, Oster, et al. (2021)
Argumentation	24	Kroll et al. (2017), Lipton (2018), and Rudin (2019)
Literature review	19	Arrieta et al. (2020), Doshi-Velez and Kim (2017), and Lepri et al. (2018)
Technical	**84**	
Method	63	Datta et al. (2016), Grabowicz et al. (2022), and J. Zhang and Bareinboim (2018)
Case study	12	Gill et al. (2020), Kung and Yu (2020), and Miron et al. (2021)
Framework	9	Ahn and Lin (2020), Hardt et al. (2021), and Sharma et al. (2020)
Behavioral	**26**	
Quantitative	20	Binns et al. (2018), Dodge et al. (2019), and John-Mathews (2022)
Qualitative	11	Dodge et al. (2019) and M. K. Lee et al. (2019)

if a *framework* is empirically evaluated on data, it also falls into this category. Lastly, *behavioral* work involves empirical examinations of human perceptions, needs, or feedback. While *quantitative* studies assess statistical differences, *qualitative* studies report verbatim or summarized statements of humans.

Through our qualitative analysis, we identified nine canonical claims made in recent literature that can be categorized into two fundamental groups: transparency as a *means* for fairness (Section 3.4.1) and transparency as a *threat* to fairness (Section 3.4.2). We also refer to the former as *desiderata* of transparency with respect to fairness, and we summarize them in Table 3.2 on page 64. The following sections cover each canonical claim in detail. We include the essential underlying intuitions behind each claim, explore varying forms of supporting evidence, and delve into the cautions and limitations raised in the relevant academic discourse.

3.4.1 Transparency as a Means for Fairness

During the analysis of the first group of claims, we realized the importance of delineating specific fairness desiderata to handle the vast number of claims. Further

Tab. 3.2.: Fairness desiderata of transparency, inferred from our structured literature review.

Desideratum	Exemplary claim
Transparency is a necessary condition for fairness	"Understanding the logic and technical inner-workings [*sic*] (i.e. semantic content) of these systems is a precondition for ensuring [...] fairness." (Leslie, 2019, p. 40)
Transparency is a sufficient condition for fairness	"The explanation of the decision process is a way to guarantee fairness to all people impacted by AI-related decision." (Ferreira & Monteiro, 2020, p. 2)
Transparency increases stakeholders' fairness perceptions	"Providing explanations for the outcome of the system increases laypeople's sense of understanding and both will eventually increase the level of perceived fairness." (Shulner-Tal et al., 2023, p. 19)
Transparency enables stakeholders to assess fairness	"These explanations are important to [...] identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected." (Gilpin et al., 2018, p. 1)
Transparency enables stakeholders to understand sources of unfairness	"AI explanations help identify potential variables that are driving the unfair outcomes." (Zhou et al., 2020, p. 1)
Transparency enables stakeholders to mitigate unfairness	"These explanations identify not only which parts of the training data are responsible for the bias but also how to reduce or 'repair' the bias." (Pradhan et al., 2022, p. 3)
Transparency enables stakeholders to certify fairness	"Using [explainable AI] systems provides the required information to justify results, particularly when unexpected decisions are made. It also ensures that there is an auditable and provable way to defend algorithmic decisions as being fair and ethical, which leads to building trust." (Adadi & Berrada, 2018, p. 52142)

examination revealed that numerous fairness notions and desiderata are inextricably connected with distinct stakeholder groups, as introduced in Section 2.3. The idea of evaluating such desiderata and capabilities from the perspective of various stakeholders is echoed in many recent studies (Arrieta et al., 2020; Ferreira & Monteiro, 2020; Hind et al., 2019; Langer & Landers, 2021; Sharma et al., 2020). Our analysis elucidates that multiple desiderata are indeed linked to distinct stakeholder groups—a connection that proves helpful in structuring the discourse.

Claim: Transparency Is a Necessary Condition for Fairness A prevailing notion in publications related to transparency is that transparency can assist in realizing a broad, unspecified concept of fairness. Such claims are frequently utilized as motivational statements in introductions, and present fairness as a universal but unspecified concept. A first line of research views transparency as a prerequisite for fairness. This perspective frequently originates from a normative standpoint, suggesting that for a decision to be deemed fair, it must be readily understandable (Alufaisan, Kantarcioglu, & Zhou, 2021). However, substantiating such strong and generic claims with empirical data poses a significant challenge. The nebulous nature of these claims implies that all stakeholders are in some capacity implicated or addressed. Nonetheless, certain theoretical support for the fundamental value of transparency is identifiable. In their analysis of various statements concerning ethical principles in AI, Floridi et al. (2018) discover a unanimous recognition of transparency as a critical element for the establishment of fair and ethical AI.

Langer, Oster, et al. (2021) propose to view the fulfillment of each desideratum as possessing two distinct facets: the *epistemic* and the *substantial* facet. The substantial facet focuses on the actual characteristics of the AI system—its fairness or trustworthiness, for instance. A desideratum, such as fairness, is deemed substantially satisfied when the system exhibits the corresponding properties to an adequate degree. The epistemic facet, on the other hand, concerns the stakeholders' ability to ascertain whether an AI system satisfies a particular desideratum, essentially evaluating if the system embodies the required properties. As such, a stakeholder is said to experience the epistemic facet of the fairness desideratum if they are equipped to assess the degree of the system's fairness. Regarding transparency mechanisms, Langer, Oster, et al. (2021) argue that they primarily serve as an epistemic enabler for fairness judgments, which may, in turn, be a basis for fulfilling the substantial facet.

Claim: Transparency Is a Sufficient Condition for Fairness The second canonical claim considers transparency even as a sufficient condition for fairness. Here, the prevailing intuition is that revealing the underlying processes of AI systems alone suffices to guarantee fairness (Ferreira & Monteiro, 2020). Other works addressing this claim approach the capabilities of transparency more cautiously (Cath, 2018). Acknowledging the essential role of transparency, Langer and Landers (2021, p. 8) underscore that "an explanation process alone does sometimes not suffice to satisfy the substantial facet of desiderata." Echoing this sentiment, Mittelstadt et al. (2016) argue that a fully comprehensible and auditable system can still result in undesired and unfair outcomes.

Claim: Transparency Increases Stakeholders' Fairness Perceptions Guided by the organizational justice framework of Colquitt (2001), several studies, such as Binns et al. (2018), have dissected fairness perceptions into *informational*, *procedural*, and *distributive* elements. Our analysis reveals that enhancing fairness perceptions often serves as the driving force behind initiatives for transparency. However, one must be cautious to ensure that this singular focus does not inadvertently neglect the need for alignment between perceptions and the actual fairness of the system (Schöffer & Kühl, 2021). The intent to cultivate positive perceptions of fairness in relation to AI systems is widely acknowledged (Ras et al., 2018), and this aspiration is typically interwoven with the objectives of fostering trust and acceptance (Papenmeier et al., 2019). Consequently, it is regularly proposed that it is beneficial for decision subjects to develop positive perspectives on the fairness of a system—an outcome which transparency is anticipated to promote (Shulner-Tal et al., 2023).

In qualitative experiments, humans have underscored their requirement for transparency and explanations in order to perceive a system as fair (Park et al., 2021). As deduced by Shin et al. (2022), these two aspects, transparency and fairness, are inextricably linked. To gauge the influence of transparency on perceived fairness, numerous experimental studies have assessed a variety of different interventions and techniques. For instance, Shulner-Tal et al. (2023) compare different types of explanations and find that all of them enhance perceptions of fairness, compared to the control group. However, research also suggests a nuanced impact of transparency on perceptions of fairness, subject to a variety of moderating factors. Apart from the *actual* system fairness and the corresponding outcomes, elements such as education and AI literacy of humans play pivotal roles. Interestingly, some findings suggest that explanations may either have no impact or even a detrimental effect on perceived fairness. For instance, Schlicker et al. (2021) do not find any evidence of transparency effects, while Binns et al. (2018) demonstrate that all tested constructs of fairness declined when participants were exposed to a single type of explanation. M. S. A. Lee (2019) interprets the observed dual impact of transparency on perceived fairness in the following way: on one side, transparency discloses model properties that might be at odds with people's fairness beliefs; conversely, transparency facilitates understanding, potentially making humans more insightful and predisposed towards fairness judgements.

Like other stakeholder groups, decision subjects are interested in making well-informed judgements about an AI system's fairness. However, taking into account limited access to information and lack of AI knowledge, it is often suggested that decision subjects should receive a specific set of information to participate in informed debates (C. Russell et al., 2017). Further, given their vulnerability and power

dynamics, an appeal process is often recommended, a theme echoed in discussions surrounding the "right to explanation" (Cath, 2018; Goodman & Flaxman, 2017). Hind et al. (2019, p. 124) underscore the value of transparency as a beneficial instrument for decision subjects to "help them understand if they were treated fairly and what factor(s) could be changed to get a different result." Addressing three transparency desiderata for decision subjects, Wachter et al. (2018) strengthen these demands by advocating for comprehensible explanations that facilitate the justification of decisions, aid in disputing them, and offer guidance on recourse. Gupta et al. (2019) conceptualize recourse as a form of explanation that offers directions on how to attain a positive outcome. Their work also illuminates how the provision of this recourse-related information can lead to unfair repercussions, thereby underscoring the concept of informational fairness and its potential inverse.

Certain empirical studies suggest that transparency may in fact help humans recognize unfairness in models and adjust their perceptions after bias mitigation (Dodge et al., 2019). Furthermore, the mere act of providing explanations can positively affect perceptions (Eiband et al., 2019), suggesting that an explanation itself might be seen as a substantial fairness value (Shin, 2021b). However, the effectiveness of transparency mechanisms may be nullified depending on the context of deployment (Binns et al., 2018). Several scholars have also voiced concerns that even accurate information may lead to skewed fairness perceptions (Gilpin et al., 2018). Angerschmid et al. (2022), for instance, suggest that perceived fairness may be heightened by transparency mechanisms, even if the AI system is fundamentally unfair. This prompts an alarming question: are humans genuinely capable of discerning the fairness of an AI system? This ongoing discourse is bolstered by ample evidence suggesting that information can be delivered in a biased manner. The central point of contention here pertains to misleading explanations, which are inherently problematic owing to their intention to deceive. Aïvodji et al. (2019) argue that the lack of specificity in transparency requirements engenders an environment that promotes the production of deceptive explanations. John-Mathews (2022) puts forth the concept of *denunciatory power*, characterizing it as the potential of an explanation to unveil instances of unfairness. Their empirical study reveals how system providers are inclined to select explanations that attract the least criticism. Yet, the propagation of misleading explanations is not limited to malicious intentions (Ehsan & Riedl, 2021). Watson and Floridi (2021) present a game-theoretic framework, identifying accuracy, simplicity, and relevance as the cornerstone characteristics of effective explanations. This suggests that the conveyed information ought to be trustworthy, easily understood, and beneficial to the recipient. They formally

demonstrate that even explanations considered accurate can lead astray if they are hard to comprehend or if their informational content lacks relevance.

Claim: Transparency Enables Stakeholders to Assess Fairness Bias detection is the initial step for most fairness-related endeavors in AI-informed decision-making. In this regard, traditional evaluation measures for prediction tasks, such as accuracy, are typically insufficient (Lipton, 2018). The resulting gap is often expected to be filled by transparency, which can measure the influence of certain protected features on outcomes, and generate statistical distribution metrics (Barocas et al., 2019). Transparency mechanisms, as Rosenfeld and Richardson (2019) suggest, allow system developers to verify the correlation between inputs and outputs, thereby confirming that fairness and other legal requirements are met. Certain bias mitigation efforts also utilize transparency for validating their success (Anders et al., 2022; Stevens et al., 2020). Moreover, some studies use transparency techniques to track fairness over time and identify any violations (Castelnovo et al., 2021; A. Ghosh et al., 2022).

Numerous empirical studies conduct fairness assessments based on the use or disuse of protected features, despite the limitations of the "fairness through unawareness" idea, as discussed in Section 2.5.4. Widely used transparency mechanisms for this task include inherent model transparency (Meng et al., 2022; Raff et al., 2018; Tolan et al., 2019), feature importance measurements based on LIME or SHAP (Alves, Bhargava, et al., 2021; Cesaro & Gagliardi Cozman, 2019; Jain et al., 2020), and counterfactual explanations (Galhotra et al., 2021; Sharma et al., 2020; Sokol & Flach, 2019). Transparency mechanisms are also incorporated into fairness testing tools, generating effective test cases (A. Aggarwal et al., 2019) and identifying problematic subsets within test data (Chung et al., 2019). However, there are several caveats and criticisms to be considered. Alikhademi et al. (2021) analyze popular transparency tools and find that merely applying post-hoc explanations does not yield sufficient insights, unless they are embedded into a broader evaluation framework, such as `AI Fairness 360` (Bellamy et al., 2019). In addition, Meng et al. (2022) echo a popular critique in fairness research, stating that simple feature importance neglects confounders and reveals nothing about causal relationships.

Claim: Transparency Enables Stakeholders to Understand Sources of Unfairness Tightly associated with fairness assessment is the analysis of sources of unfairness. This task often encompasses elements of fairness assessment but advances further to elucidate particular drivers. Echoing Langer, Oster, et al. (2021), this task

enhances the observational epistemic facet with an analytical aspect. In this regard, transparency is to serve not only the observation of unfairness but to unearth the underlying mechanisms responsible for this observation. A common method in the literature is to measure the importance of protected attributes, aiming to nullify their influence. By this logic, observing the use of protected attributes simultaneously uncovers the cause of unfairness—if we define *unfairness* as the use of such protected features. Several intuitive claims support the distinction between assessment and analysis (Abdollahi & Nasraoui, 2018; Arrieta et al., 2020; Leslie, 2019). A survey by Zhou et al. (2020) further claims that transparency aids in identifying the features that drive statistical notions of unfairness.

A plethora of studies focus on investigating and comprehending the emergence of unfairness in AI systems. Pradhan et al. (2022), for instance, devise a data-based explanation technique assuming that many transparency frameworks report bias but fail to pinpoint its origins. Advocating a more rigorous feature engineering procedure, Siering (2022) employ information diagnosticity theory to understand how certain features could potentially induce unfairness for relevant stakeholders. Ahn and Lin (2020) propose a framework and visual analytic system guiding the complete machine learning pipeline to understand impacts on various fairness metrics. The most prevalent type of research, however, centers around post-hoc transparency. As early initiators, Datta et al. (2016) consider correlations when quantifying the influence of features. Advancing and refining preexisting Shapley frameworks, Begley et al. (2020) and Miroshnikov et al. (2022) calculate the impact of features on fairness metrics. Moreover, B. Ghosh et al. (2023) address issues related to intersectional fairness (also referred to as *fairness gerrymandering*), which come into play when multiple attributes such as gender and race warrant protection. This type of analysis is often paired with efforts to mitigate bias, for instance, by scrutinizing feature usage both prior to and following the implementation of bias mitigation strategies, as detailed by Grabowicz et al. (2022). In line with this, Quadrianto et al. (2019) strive not just to determine *whether* but also *how* unfairness is achieved in the context of representational learning. General criticism regarding the reliability of transparency mechanisms applies to these claims as well. We cover such critique in more depth in Section 3.4.2. Particular reference can be made to Alikhademi et al. (2021), who discover deficiencies for several popular transparency methods, not only in bias assessment but also in analyzing sources of unfairness. More broadly, Warner and Sloan (2021) posit that there are fairness aspects that transparency will never be capable of uncovering, for instance, taking into account adversarial circumstances leading to unfairness within the training data.

Claim: Transparency Enables Stakeholders to Mitigate Unfairness Previously discussed claims hint at an evident symbiotic relationship between transparency and bias mitigation. In fact, numerous empirical studies on transparency suggest that their bias analysis is a solid foundation for subsequent bias mitigation steps (Tolan et al., 2019). A prevalent belief is that awareness and understanding of unfairness heighten the probability of its mitigation (Meng et al., 2022). This aligns with the distinction between epistemic and substantial aspects of fairness (Langer, Oster, et al., 2021), and claims suggesting that unintentional bias can be alleviated if detected (Franke, 2022). Numerous empirical studies incorporate transparency methods into their bias mitigation techniques, either as a part of a fairness-enhancing algorithm or as interpretable fairness constraints. Some pre-processing techniques employ causal explanations to identify and resolve unfair patterns in the training data (Pradhan et al., 2022). However, the majority of prior work focuses mostly on retraining the model, that is, post-processing methods. For instance, Hickey et al. (2021) retrain their model using a fairness regularization term calculated using SHAP. Similarly, Dash et al. (2022) construct counterfactual explanations and utilize them as a regularization technique to reduce bias. To address unfairness emerging from concept drift, A. Ghosh et al. (2022) establish a monitoring system that they claim to be capable of automatically detecting and mitigating bias based on a novel feature importance quantification.

Another way transparency may directly contribute to fairness is by creating better trade-offs between fairness and accuracy. As accuracy is often a central goal for system developers, it is vital that the "cost of fairness" (von Zahn et al., 2022) is not overly burdensome. Ge et al. (2022), for instance, propose a transparency mechanism to extract features for a recommendation system that enhance this trade-off against several benchmarks. The trade-offs between transparency, fairness, and accuracy are further discussed in Section 3.4.2. A relevant caveat is noted by Karimi et al. (2022), who formally demonstrate that bias mitigation techniques relying on feature importance overlook the fairness of recourse. Consequently, fairness criteria should extend beyond distributional metrics like demographic parity, and consider the actions necessary to receive a positive outcome (Gupta et al., 2019). Finally, Waller and Waller (2022) contend that bias mitigation in AI systems is confined to predefined protected groups, but concurrently generates a so-called *assembled bias* that may detrimentally affect unforeseen categories of people beyond our notice.

Moving the spotlight towards the actual users of AI systems, it becomes apparent that many advocate for transparency to empower an informed human-in-the-loop. For instance, Ahn and Lin (2020) contend that in the absence of generally applicable criteria, transparency enables the incorporation of domain knowledge to realize

informed trade-off decisions between fairness and utility. In other words, decision makers should be enabled to determine the extent of predictive performance that ought to be sacrificed to attain a certain "degree" of fairness. This form of human-AI collaboration requires reliable transparency mechanisms that users can trust and utilize to counteract unfair decisions. Generally, it is often assumed that through simplification and visualization, even non-technical users can comprehend AI systems, allowing them to apply their human judgement and domain knowledge. This intuition is echoed in a range of conceptual works such as by Wagner and d'Avila Garcez (2021), who aim to iteratively incorporate user knowledge based on evaluations with transparency mechanisms. Stumpf et al. (2021) underscore decision makers as a crucial stakeholder group whose needs should be explicitly catered to when designing transparency mechanisms for fairness. They further validate this by conducting co-design workshops with domain experts. Park et al. (2022) follow a similar path, carrying out comprehensive workshops in the field of human resources management, and discover that managers require easily understandable, global explanations to endorse AI-informed decisions.

Beyond that, several technical works acknowledge the need for user input in their empirical evaluations. For instance, by leveraging interpretable rule lists, Aïvodji et al. (2021) calculate and visualize several accuracy-fairness trade-off curves to facilitate analyses by human domain experts. Transparency has also been suggested to equip users to directly mitigate model bias. For instance, Y. Zhang and Ramesh (2020) demonstrate that humans without extensive AI knowledge can directly integrate domain-specific interpretable constraints into the model. Similarly, Chakraborty et al. (2020, p. 3) propose a transparency method that illustrates the nearest neighbors of an unfairly classified data point and permits users to "easily evaluate our explanations and take decision whether to change the prediction or not." Together, these narratives sketch a picture of an environment where transparency not only helps reduce bias and improve fairness, but also empowers humans, particularly decision makers, to actively contribute to these processes. The power of visualization, simplification, and domain-specific knowledge comes into play in these settings, with human-in-the-loop strategies offering a viable path towards improved model fairness and utility. However, the successful implementation of such strategies is contingent upon the development and use of trustworthy, reliable, and understandable tools and methods.

Claim: Transparency Enables Stakeholders to Certify Fairness From system deployers' perspectives, Cornacchia et al. (2021) argue that loan providers should employ explanations to ensure fair decisions for their clientele. Hind et al. (2019)

identify regulators as a vital stakeholder group, who intend to guide algorithmic decisions towards socially accepted norms. According to Ras et al. (2018), the responsible use of AI systems requires regulatory structures that extend beyond the capabilities of existing transparency mechanisms, yet, transparency is frequently suggested as a beneficial tool for auditing purposes (Adadi & Berrada, 2018). Following this line of reasoning, system deployers are required to demonstrate to external auditors that their system is ethically sound, non-discriminatory, and worthy of public trust and safety, as noted by Leslie (2019). While these audits are often associated with broader terms like accountability or software testing (Zucker & d'Leeuwen, 2020), transparency mechanisms are expected to yield insights into AI systems themselves. Numerous conceptual works advocate for specific forms of auditability or other transparency requirements. In addition to external auditing, Floridi et al. (2018) suggest the necessity for internal auditing mechanisms, enabling system deployers to take more responsibility. Loi et al. (2021) advocate for a form of transparency, termed "design publicity," that not only explains and justifies a decision at hand but also communicates the underlying goals and values.

Several papers distinguish between transparency and auditability, emphasizing their different objectives. Springer and Whittaker (2019) suggest that while transparency focuses on comprehension and user experience, auditability should enable auditors to scrutinize model fairness. Similarly, Warner and Sloan (2021) propose transparency for a regulatory purpose, aligning with the concept of auditability. Shulner-Tal et al. (2023) advocate for the implementation of audit-based certifications, conceptualizing them as a form of explanation intended for stakeholders impacted by AI systems. The purpose of such certifications is to reconcile statistical notions of fairness—discernible in the system's auditing process—with human perceptions of fairness. Meanwhile, Gryz and Shahbazi (2020) assert that the demand for explanations and direct feedback from decision subjects predominantly stems from the nascent nature of AI systems. They further argue that, analogous to protocols followed in other high-stakes sectors such as bridge safety, society ought to entrust expert auditors with the responsibility of verifying the fairness of AI systems.

Several technical works claim their practicality for auditing. For instance, Sharma et al. (2020) develop a framework that equips regulators to inspect the robustness, fairness, and transparency of a black-box model. Hickey et al. (2021) put forth and test a novel fairness definition aimed at external auditors seeking to scrutinize a surrogate model for statistical notions of fairness. Based on an experimental study, John-Mathews (2022) outlines two scenarios where transparency mechanisms can assist in realizing ethical principles. The first scenario calls for a strict formalization of transparency requirements that auditors can use to evaluate an AI system's

compliance with fairness objectives. The second, more liberal, scenario offers more degrees of freedom to system deployers but necessitates standardized, randomized experiments that allow decision subjects to disclose unfair decisions.

The distinction between transparency and auditability has prompted some scholars to posit that transparency is not an absolute necessity for ensuring fairness. Works by Springer and Whittaker (2019) and Warner and Sloan (2021) advocate that statistical analyses performed on test data are ample for the assessment of statistical notions of fairness. A parallel discourse is ongoing around the necessity of source code transparency, with viewpoints suggesting that it is not an essential nor a sufficient prerequisite for fairness certification (G. K. Y. Chan, 2022). Moreover, despite the often assumed usefulness of local explanations for decision subjects, it has been repeatedly observed that these explanations hold limited value for auditors, who primarily focus on evaluating the fairness of the entire AI system (Wachter et al., 2018). Conversely, global explanations are considered to hold promise as an auditing tool, despite the need to be "simplified to the point of absurdity in order to be intelligible," as stated by Seymour (2018, p. 4). Lastly, Loi et al. (2021) stress that feature-based explanations are inadequate in providing a normative justification for the fairness of using such features.

3.4.2 Transparency as a Threat to Fairness

The following canonical claims reflect common critique of transparency mechanisms with respect to fairness. Consequently, the discussions and implications touch on and influence all stakeholders involved. For instance, if an explanation is fundamentally misleading, it results in ripple effects on system developers, leading to inaccurate fairness assessments; on regulators, causing an inability to certify fairness; and on decision subjects, leaving them without the means to judge if they are treated fairly. Likewise, the discourse around transparency-fairness trade-offs holds relevance for multiple stakeholder groups.

Claim: Transparency Mechanisms Are Prone to Misinterpretation and Manipulation
Initiating a comprehensive debate on the faithfulness and persuasiveness of transparency mechanisms, Herman (2017) raises the question of how much simplification can be ethically justified for persuading human stakeholders. Gilpin et al. (2018) address this dilemma by advocating for interventions that balance transparency and completeness. In a compelling call for inherently interpretable models, Rudin (2019, p. 207) argues that post-hoc explanations for black-box models must be

misleading. Concretely, she argues that "even an explanation model that predicts almost identically to a black box model might use completely different features." Furthermore, she cautions that there is no guarantee that an accurate explanatory surrogate model enhances understanding of the black-box model it seeks to explain. Galinkin (2022) brings attention to the potential for transparency mechanisms to be purposefully skewed to give an illusion of fairness. Therefore, Gill et al. (2020) propose resilience against deceptive explanations as a primary design criterion for AI systems.

The concept of manipulation is both theoretically and empirically introduced by Aïvodji et al. (2019) under the term *fairwashing*. They demonstrate how an unfair model can be conveniently transformed into a model of identical performance, whose feature importance scores flawlessly satisfy any chosen fairness metrics. Similarly, Dimanov et al. (2020) illustrate how unfair models can mask the use of protected features from six commonly used transparency mechanisms. Anders et al. (2020) show how feature importance scores can be manipulated arbitrarily without impacting the model performance. In the same vein, Slack, Hilgard, et al. (2020) reveal limitations of LIME and SHAP by exploiting the perturbations generated by these techniques to control post-hoc explanations. In conclusion, all popular post-hoc explanations based on feature importance have been proven susceptible to manipulations. Therefore, Begley et al. (2020) conclude that feature importance based on protected attributes is inadequate for measuring fairness.

Balagopalan et al. (2022) bring to light another risk posed by inadequate transparency mechanisms. They suggest that varying quality of explanations (referred to as *fidelity gaps*) represents a novel type of unfairness that can harm protected groups, especially in terms of harmful downstream decision-making consequences. This observation is validated by Dai et al. (2022), who identify disparities in fidelity, consistency, stability, and sparsity across several transparency methods. Another line of criticism focuses on the constraints imposed by human fallibility. Lipton (2018, p. 22) notes that individuals often choose explanations that align with their subjective interests, making them susceptible to "misleading but plausible" explanations. This is further reinforced by Selbst and Barocas (2018) and Walmsley (2021), who discuss an inherent disconnect between human intuition and the statistical patterns detected by AI systems, complicating any normative evaluation. Herman (2017) argues that cognitive limitations in interpreting explanations necessitate simplification, and this cognitive bias fundamentally jeopardizes the pursuit for ethical AI. In response to these human limitations, efforts have been made to enhance the robustness of explanations, navigating the complexities inherent in this endeavor. Such endeavors have been undertaken, for instance, by Aïvodji et al. (2021), Anders et al. (2020),

Begley et al. (2020), and Chakraborty et al. (2020). However, it is worth noting that the term *robustness* is commonly employed in this context, but a universally accepted definition remains elusive. For instance, Sharma et al. (2020) suggest a robustness score that measures consistency of explanations amid minor changes in input or hyper-parameters. On the other hand, Gill et al. (2020) consider robustness as a defense mechanism against adversarial attacks, particularly in relation to manipulation.

Claim: Transparency and Fairness Are Conflicting Goals Adomavicius and Yang (2022) argue that transparency can at times contribute to fairness and at other times hinder it. Padmanabhan et al. (2020) consider fairness and interpretability as equally crucial design goals, but they acknowledge that achieving both simultaneously can pose challenges. This is corroborated in a case study by C. Wang et al. (2022), who note that most techniques for mitigating bias depend on non-transparent transformations. Furthermore, several conceptual works assert that insights gained through transparency could potentially confer unfair advantages to actors who exploit AI systems (Gryz & Shahbazi, 2020; Park et al., 2022). Regarding the statistical relationship between transparency and fairness, Jabbari et al. (2020) find that the impact of transparency on fairness metrics hinges on the predictive value of protected attributes and the separability of classes. As complexity increases, accuracy follows a monotonic growth, whereas fairness metrics trace different trajectories with varying positive and negative effects.

Numerous papers argue that the discussed trade-off actually involves three variables, with accuracy emerging as the third key factor. Kleinberg and Mullainathan (2019, p. 1) formally demonstrate that transparency incurs a cost in terms of lower accuracy and fairness, stating that "using a simple prediction function both reduces utility for disadvantaged groups and reduces overall welfare relative to other options." Furthermore, they find that simpler models increase the incentive to incorporate protected attributes due to their predictive value. Recognizing that both transparency and fairness compromise accuracy (and hence economic value), Borrellas and Unceta (2021) contend that these relationships ultimately culminate in a social welfare trade-off steered by economic interests, legal constraints, and public scrutiny. Adopting a similar perspective, Selbst and Barocas (2018) argue that transparency and fairness are entwined in a broader trade-off among additional normative objectives, including privacy. We summarize all canonical claims including references to prior work that addresses them in Tables A.1 and A.2 in Appendix A. Note that some papers address more than one of the claims.

3.5 Conclusion

Upon analyzing 169 papers that discuss the interplay of transparency and fairness, our conclusion reveals the role of transparency to be threefold: (i) it is multidimensional, with diverse stakeholders each pursuing different fairness desiderata; (ii) it is nuanced, given that transparency has been demonstrated to be effective only for specific fairness desiderata; and (iii) it is multi-modal, as transparency can influence these fairness desiderata in a variety of ways. Moreover, we discover that many arguments lack either empirical evidence or comprehensive argumentative reasoning, thereby raising the concern that such claims could foster undue optimism regarding the present use of transparency mechanisms. These insights form an important cornerstone for the remainder of this thesis.

A large proportion of the literature we surveyed supports transparency for various fairness desiderata. Most studies underscore the epistemic capacity of transparency, which is sometimes able to empower human stakeholders to evaluate certain fairness criteria and even to analyze sources of bias. In some instances, transparency mechanisms are so thoroughly integrated into bias mitigation techniques or workflows that they can directly impact substantial concepts of fairness. However, we caution that recognizing unfairness does not necessarily lead to fairness, and we endorse the distinction between epistemic and substantial facets of fairness desiderata, as proposed by Langer, Oster, et al. (2021). Regarding regulatory purposes, the role of transparency is highly contentious, largely due to the uncertain legal landscape. In contributing to the discussion on auditability, some researchers propose that transparency can be invaluable for certifying procedural fairness (i.e., model reasoning), but for certifying specific notions of distributive fairness (i.e., model output), simple fairness testing based on test data may be adequate. It is also important to underline the complex, contested nature of fairness, as evidenced by numerous formal impossibility results (Chouldechova, 2017; Kleinberg et al., 2017). This indicates that one may never be able to confer a universally acceptable certification of fairness.

Our work also offers a thorough overview of both empirical and conceptual criticisms of transparency. While transparency may sometimes be useful for measuring feature importance, we caution that such measures should be approached with skepticism when assessing various notions of fairness. Not only are feature importance values susceptible to manipulation, but recent studies also reveal their failure to account for correlations with proxy variables. To address deficiencies of existing metrics for fairness perceptions, our work suggests the need for a fresh conceptualization of

informational fairness. This should account for the demands of decision subjects for truthful and helpful information regarding the underlying normative setting, the AI system including its outcomes, and potential responses to those outcomes. While transparency mechanisms may provide some of this information, we recognize that transparency might not always be necessary for all fairness dimensions—and it is certainly not sufficient. Finally, we echo previous calls for a sociotechnical perspective, asserting that transparency is only one of many considerations for achieving an ethical objective as intricate and diverse as fairness.

Part II

Fully Automated Decision-Making

4 Designing Inherently Transparent and Fair AI Systems

In this chapter, we construct an artificial intelligence (AI) system that upholds both inherent model transparency and common statistical fairness notions. In doing so, we demonstrate that transparency and fairness are not necessarily mutually exclusive objectives. Our methodology employs a ranking-based approach rooted in monotonic relationships between legitimate features and the outcome, a factor often deemed crucial for inherent transparency in AI systems. This approach is predicated on a distance-based decision criterion that leverages legitimate information from historical data and addresses problematic correlations between protected and (seemingly) legitimate features. Through a comprehensive series of experiments, we illustrate that our methodology outperforms traditional supervised machine learning (ML) methods on a range of relevant fairness metrics—especially in the presence of strong label bias.

4.1 Introduction

AI systems have been increasingly used for decision support in recent years. A common perception is that algorithms can avoid human bias and make more objective and transparent decisions (Castelluccia & Le Métayer, 2019). However, as algorithms support humans with evermore consequential decisions, they have also become subject to enhanced scrutiny. In 2016, journalists at ProPublica found that

This chapter is based on published work. To enhance the reading experience and maintain overall consistency of the thesis, we removed the abstract and made several minor adjustments. The original paper can be accessed via:

Schöffer, J., Kühl, N. & Valera, I. (2021). A ranking approach to fair classification. *COMPASS '21: ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 115–125). https://doi.org/10.1145/3460112.3471950.

COMPAS, a system used by US courts to assess defendants' risk of recidivism, was unfair towards Black people (Angwin et al., 2016). In November 2019, Bloomberg reported that Steve Wozniak was suspecting the algorithm that determines credit limits for Apple's credit card to discriminate against women (Nasiripour & Natarajan, 2019). These and other examples make obvious the need for understanding root causes and developing techniques to combat algorithmic unfairness. In large part, prior work has focused on formalizing the concept of fairness and enforcing certain statistical equity constraints when making predictions—mostly in a setting of binary classification, for instance, when it must be decided whether a loan should be offered or not.

However, traditional classification algorithms require access to *actual* ground-truth labels, which are often unavailable (Lakkaraju et al., 2017). In practice, we may only have access to *imperfect* labels (Rädsch et al., 2021), often as the result of (potentially biased) historical human-made decisions. Inspired by the argumentation of Kilbertus et al. (2020), we propose to *not* learn to predict imperfect labels. Instead, we introduce a meritocratically fair decision criterion based on an observation's distance to what we call the *North Star*—a (potentially hypothetical) observation that is most qualified in a given scenario. Our approach induces both a ranking and an opportunity to classify observations, based on monotonic relationships between legitimate features and the outcome. We also put forward ideas to (i) incorporate useful information from historical decisions (Section 4.3.2) and (ii) reduce the importance of features that are highly correlated with protected attributes (e.g., gender) in the decision-making process (Section 4.3.3).

The rest of the work is structured as follows: in Section 4.2, we introduce important concepts and related work. Section 4.3 represents the core of our work—the methodology as well as theoretical results. In Section 4.4, we illustrate our method by the example of the German Credit dataset, and we conduct extensive experiments on synthetic data in Section 4.5. Section 4.6 summarizes our work, discusses its limitations, and provides potential areas for follow-up work.

4.2 Background

To lay out the foundations, we briefly introduce important concepts related to our proposed methodology. We start with a summary of the notation used in this work. We call A the set of *protected* features which must not be discriminated against, and let $a_k \in A,\ k \in \{1, \ldots, K\}$ be the individual protected features. The US Equality

Act (United States Congress, 2019), for instance, defines sex, gender identity, and sexual orientation, among others, as protected features. In line with other related work, we assume that the decision whether a feature is protected or not is made externally (Žliobaitė, 2015). We further define $x_\ell \in X,\ \ell \in \{1, \ldots, L\}$ as the non-protected (or *legitimate*) features, and Y as the set of *imperfect* labels, which can be either positive $(+)$ or negative $(-)$. We call these labels *imperfect* as they are only a noisy signal of the true labels. A given dataset consisting of A, X, Y is referred to as $\mathcal{D}$. Lastly, we call $\widehat{Y}$ the predictor, a function that maps observations to positive or negative outcomes. When referring to an individual observation, we use superscripts like $A^{(i)}$, which would be the set of protected features for observation (i), while we have N observations in total.

4.2.1 Relevant Notions of Fairness

Mehrabi et al. (2021, p. 11) define fairness in the context of decision-making as the "absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits." Generally, existing literature distinguishes individual from group fairness definitions. In this work, we are primarily concerned with individual fairness—however we also make group fairness-related arguments in the spirit of demographic parity (Zafar et al., 2019) later on. A typical approach aiming at individual fairness is *fairness through awareness* (FTA) (Dwork et al., 2012). We will briefly introduce this, as well as the conception of "fairness through unawareness" (FTU) (Grgić-Hlača et al., 2020), due to their importance for our work.

In line with Grgić-Hlača et al. (2020), FTU is the conception that an algorithm is fair so long as any protected features are not explicitly used in the decision-making process. In other words, FTU simply requires a predictor $\widehat{Y}$ to ignore all protected features in A. However, as Hardt et al. (2016) argue, this definition is ineffective in the presence of strong correlation between protected and legitimate features. In the work at hand, we address this issue by penalizing highly correlated features with respect to their importance for decision-making.

According to Dwork et al. (2012), FTA says that an algorithm is fair if it gives similar predictions to similar individuals. Formally, FTA requires an appropriate distance metric $d(\cdot, \cdot)$. If, for two individuals (i) and (j), $d(i, j)$ is small, then FTA requires that $\widehat{Y}^{(i)} \approx \widehat{Y}^{(j)}$. As stated by Dwork et al. (2012), the main challenge with this notion is defining an appropriate distance metric. In most cases, this requires domain-specific knowledge.

4.2.2 Related Work

Most of existing work on algorithmic fairness has been concerned with fair classification. Herein, numerous articles have been published on how to formally define (Corbett-Davies et al., 2017; Dwork et al., 2012; Grgić-Hlača et al., 2020; Hardt et al., 2016; Kusner et al., 2017; Pedreshi et al., 2008) and enforce (Calmon et al., 2017; Hardt et al., 2016; Kamiran & Calders, 2009, 2012; Kamishima et al., 2012; Kilbertus et al., 2020; Zafar et al., 2019) fairness. Generally, fairness-aware techniques can be divided into three categories which are based on the process step of the application: the first category is concerned with removing existing bias from the training data (pre-processing). Typical approaches involve transformation of the data (Calmon et al., 2017) or changing the class labels for training (Kamiran & Calders, 2012). The second category involves modification of existing algorithms (in-processing), typically through adding fairness constraints (Zafar et al., 2019) or penalizing discrimination, for instance, by means of regularization (Kamishima et al., 2012). The third category includes all techniques aimed at changing the output of a potentially unfair model (post-processing). Hardt et al. (2016), for instance, construct a non-discriminating predictor from an existing one via solving an optimization problem. However, if ground-truth labels are not (or selectively) available (Lakkaraju et al., 2017), then maximizing for prediction accuracy seems counter-intuitive and is, in fact, sub-optimal (Kilbertus et al., 2020).

More recently, alternative concepts based on the theory of causal inference have evolved (Kilbertus et al., 2017; Kusner et al., 2017). While these approaches have shown promising results, they generally make strong assumptions about the causal structure of the world. An in-depth discussion (including common misconceptions) of causal models in the realm of algorithmic fairness is provided, for instance, by L. Hu and Kohler-Hausmann (2020).

Fair ranking approaches can be split into pre-processing (K. Yang & Stoyanovich, 2017), in-processing (Zehlike & Castillo, 2020), and post-processing (Biega et al., 2018; Celis et al., 2018; Singh & Joachims, 2017, 2018; Zehlike et al., 2017) techniques as well (Castillo, 2019). With respect to quantifying fairness, most of existing methods apply an attention-based criterion, aiming at equalizing exposure of observations in, for instance, (web) searches (Biega et al., 2018; Singh & Joachims, 2017, 2018). Furthermore, a majority of existing literature has been focusing on achieving group fairness, whereas individual fairness considerations for rankings remain scarce—with few exceptions, such as work by Biega et al. (2018), where the authors introduce a mechanism to achieve individual fairness across a *series* of

rankings. To the best of our knowledge, no existing work on fair ranking is closely related to ours in terms of methodology.

Perhaps the most related work is an article by S. Wang and Gupta (2020). Here, the authors put forward the idea of optimizing classification accuracy subject to a classifier being monotonic in a given set of features. Thereby, it is argued, the classifier can evade violating "common deontological ethical principles and social norms such as [...] 'do not penalize good attributes'" (S. Wang & Gupta, 2020, p. 1). While the idea of enforcing monotonicity constraints is similar, we have identified three major differences to our work:

1. S. Wang and Gupta (2020) use supervised ML to predict ground-truth labels, whereas we assume imperfect labels. Our proposal in the case of imperfect labels is to *not* maximize for accuracy in the first place.
2. They do not take measures to prevent the algorithm from "exploiting" protected information to achieve higher accuracy.
3. They do not account for the well-known problem of indirect discrimination, which occurs when (seemingly) legitimate features are highly correlated with protected features.

4.3 Proposed Methodology

In this chapter, we introduce our proposed ranking algorithm for decision-making with imperfectly labeled data, that is, the common case where ground-truth labels are not available. Specifically, we assume we are given data $\mathcal{D}$ with imperfect labels stemming from human-made decisions, for instance, whether an applicant was admitted to graduate school or not.

Our approach follows a notion of individual fairness that aims at uniting both fairness definitions from Section 4.2.1, FTU and FTA. Note that this idea is closely related to the concept of "meritocratic fairness," as coined by Kearns et al. (2017). We call an algorithm *meritocratically fair* if it assigns the positive outcome to the most qualified observations, regardless of protected features. This definition is in line with many equal employment opportunity policies, yet disregarding affirmative action. Based on this notion, we can also define meritocratic *un*fairness:

Definition 4.1 (Meritocratic unfairness)**.** *An individual observation* (i) *is treated unfairly over a different observation* (j) *if* (i) *is more qualified than* (j) *but: a) ranked lower, or b) assigned* $(-)$ *while* (j) *is assigned* $(+)$*.*

We will use Definition 4.1 for evaluation purposes later on. However, a definition of what *qualified* means can hardly be given without knowledge of the respective use case—a viewpoint that is shared, among others, by Dwork et al. (2012). We will address this now.

To illustrate our ideas, we construct a (simplified) synthetic graduate school admission dataset and use it as a running example—an excerpt is shown in Table 4.1 on page 87. The dataset consists of 1000 observations, with 50% being males and 50% females (protected feature). The Graduate Record Examination (GRE) scores (legitimate features) are three-dimensional: GRE V(erbal Reasoning), GRE Q(uantitative Reasoning), GRE A(nalytical) W(riting). We sample them from multivariate Gaussian distributions

$$\mathcal{N}(\mu_{m(ale)}, \Sigma_m), \quad \mathcal{N}(\mu_{f(emale)}, \Sigma_f),$$

where

$$\mu_m = [150.7, 156.1, 3.5], \quad \mu_f = [150.3, 151.2, 3.7],$$

and the covariance matrices

$$\Sigma_m = \begin{bmatrix} 81.00 & 28.15 & 5.43 \\ 28.15 & 84.64 & 1.16 \\ 5.43 & 1.16 & 0.81 \end{bmatrix}, \quad \Sigma_f = \begin{bmatrix} 65.61 & 24.51 & 4.34 \\ 24.51 & 79.21 & 1.00 \\ 4.34 & 1.00 & 0.64 \end{bmatrix}$$

are derived from the official data provided by the administrator of the GRE test (ETS, 2019, 2018). For compliance with the official ranges of scores, we round and truncate the sampled scores such that a) GRE V and GRE Q scores are between 130 and 170 in one-point increments, and b) GRE AW scores are between 0 and 6 in half-point increments. To simulate historical admission decisions, we scale the legitimate features between 0 and 1, and generate imperfect labels as

$$Y = \begin{cases} (+) & \textit{if } \ 0.1 \cdot \mathbb{1}_{male} + 0.2 \cdot \text{GRE V} + 0.5 \cdot \text{GRE Q} + 0.2 \cdot \text{GRE AW} + \epsilon > 0.5 \\ (-) & \textit{otherwise}, \end{cases}$$

where $\mathbb{1}$ is the indicator function and $\epsilon \sim \mathcal{U}(0, 0.1)$ is noise. Note that male applicants are given an unfair advantage over their female counterparts. Apart from this, the

Tab. 4.1.: Exemplary graduate school admission data.

ID	Gender	GRE V	GRE Q	GRE AW	Y
1	*male*	147	144	3.0	$(+)$
2	*male*	146	140	3.5	$(+)$
...	...	...	...	...	...
11	*female*	153	147	3.5	$(-)$
...	...	...	...	...	...

Algorithm 1: Scaling of legitimate features.

Input : Legitimate features $x_1, \ldots, x_L$ of $\mathcal{D}$, including $(\uparrow)$ or $(\downarrow)$ relationships.
Output : Scaled features $z_1, \ldots, z_L$.
for $\ell \in \{1, \ldots, L\}$ **do**
  $z_\ell^{(i)} \leftarrow \frac{\left(x_\ell^{(i)} - \min_j x_\ell^{(j)}\right)}{\left(\max_j x_\ell^{(j)} - \min_j x_\ell^{(j)}\right)}$ for all $i \in \{1, \ldots, N\}$;
  if x_ℓ *is* $(\downarrow)$ **then**
    $z_\ell \leftarrow (1 - z_\ell)$;
  end
end

high importance of GRE Q scores could be representative of a technical university's admission process.

We assume that for any specific use case, we are given (e.g., by an expert) or can easily derive the information of how any legitimate and relevant feature should impact the final decision—specifically, whether higher $(\uparrow)$ or lower $(\downarrow)$ values are beneficial with respect to the positive outcome. Note that if certain feature interactions have a known monotonic relationship with the outcome, then these interactions can be added as additional features and assigned a $(\uparrow)$ or $(\downarrow)$ as well. Certainly, in many cases these dependencies are obvious and need not be verified by an expert. For instance, in our graduate school example, it is clear that high GRE scores are more beneficial towards being admitted than low scores. Alternatively, if obtaining this information from an expert is too expensive, we could potentially infer the $(\uparrow)$ or $(\downarrow)$ relationships from the data $\mathcal{D}$. The idea that relevant features should have a monotonic relationship with the outcome is, for instance, similarly introduced by S. Wang and Gupta (2020).

With this information, we first scale the legitimate features X such that all values are in $[0, 1]$. We call the scaled legitimate features $z_\ell \in Z,\ \ell \in \{1, \ldots, L\}$. We further require that the probability of the positive outcome increases with the value of any z_ℓ. For that, we perform $z_\ell \leftarrow (1 - z_\ell)$ if the original relationship between x_ℓ and the

outcome is $(\downarrow)$. These steps are summarized in Algorithm 1 on page 87. Note that we can also apply Algorithm 1 on page 87 to observations that are not contained in $\mathcal{D}$. In that case, we need to assume that the resulting values of z are capped at 0 and 1.

4.3.1 Measuring Distance to the North Star

Our idea is to fairly rank observations based on their distance to what we call the *North Star*.

Definition 4.2 (North Star)**.** *Given a dataset $\mathcal{D}$ and the respective legitimate features $z_\ell,\ \ell \in \{1,\dots,L\}$ scaled as in Algorithm 1 on page 87, the* North Star *is a (potentially hypothetical) observation $(\star)$ that attains the maximum observed value for each legitimate feature:*

$$z_\ell^{(\star)} \coloneqq \max_{i \in \{1,\dots,N\}} z_\ell^{(i)} = 1 \quad \forall \ell \in \{1,\dots,L\}. \tag{4.1}$$

Now, we can compute the distance of single observations to the North Star. For that, we choose the taxicab metric for its clear interpretation and its favorable behavior in higher dimensions (C. C. Aggarwal et al., 2001). Note that the approach also works for other metrics. We define the distance of an observation (i) to the North Star as follows:

$$d(i,\star) \coloneqq \sum_{\ell=1}^{L} \left(1 - z_\ell^{(i)}\right), \tag{4.2}$$

considering that $z_\ell^{(i)} \in [0,1]$ for all ℓ and observations (i). Note that we assume symmetry of our distance measure, that is, $d(i,\star) = d(\star,i)$. In our example, this distance would be 0 for applicants with the perfect scores of GRE V $= 170$, GRE Q $= 170$, and GRE AW $= 6.0$.

4.3.2 Extracting Useful Information From Historical Decisions

In a next step, we enhance the distance formula in Equation (4.2) with useful information from historical data. Despite the fact that our given data $\mathcal{D}$ contains only imperfect labels, we argue that historical decisions often contain *some* useful information on the true labels. Specifically, we aim to extract the relative importance of legitimate features from historical decisions, assuming that important features

from the past are still important at present. In our running example, for instance, we know that labels are biased, but we still want to capture that GRE Q scores are most important for admission at a technical university.

Our rationale is the following: while Equation (4.2) implicitly treats every feature as having equal importance, we want to account for the fact that some features are undoubtedly more important than others for decision-making. Even though we could explicitly ask experts for this information, similar to the $(\uparrow)$ or $(\downarrow)$ relationships, we argue that manually quantifying the importance of individual features (in %) is often intractable. We, therefore, propose to learn these importances directly from $\mathcal{D}$, for instance, through the concept of permutation importance (Breiman, 2001), which is defined as the decrease in model score when the value of this respective feature is randomly permuted. The result is an estimate of how much a given model depends on this feature. For that, we train a classifier on $\mathcal{D}$ and obtain feature importances $\omega_1, \ldots, \omega_L$, with $\omega_1, \ldots, \omega_L \geq 0$ and $\sum_{\ell=1}^{L} \omega_\ell = 1$ for all legitimate features.[1] In a next step, we can now adjust Equation (4.2) by adding $\omega_1, \ldots, \omega_L$ as weights to reflect the importance of each legitimate feature:

$$d'(i, \star) \coloneqq \sum_{\ell=1}^{L} \omega_\ell \left(1 - z_\ell^{(i)}\right). \tag{4.3}$$

In the running example, we obtain $\omega_V = 0.10$, $\omega_Q = 0.73$, and $\omega_{AW} = 0.17$, with standard deviations $\sigma_V = 0.006$, $\sigma_Q = 0.012$, and $\sigma_{AW} = 0.005$, by fitting a random forest classifier with bootstrapping and using the `permutation_importance` function of `scikit-learn` (Pedregosa et al., 2011). This means that d' would be most sensitive to changes in GRE Q scores, as desired.

4.3.3 Accounting For Relationships Between Legitimate and Protected Features

As outlined in Section 4.2.1, as well as by Hardt et al. (2016) and Pedreshi et al. (2008), the fundamental weakness of FTU as a notion of fairness is the fact that protected features can sometimes be predicted from legitimate features. It is particularly problematic if legitimate features are highly correlated with protected features—we account for this by penalizing high correlation.

[1] It might happen that the legitimate features cannot predict the historical labels reasonably well (e.g., if labels are random). In such cases we can skip this step.

To measure general monotonic (i.e., not just linear, as Pearson's r would do) relationships between two data samples, we can use Spearman's rank correlation coefficient (SRCC). For the rankings rk_a and rk_x of two samples a and x, SRCC $\rho_{a,x}$ is calculated as follows:

$$\rho_{a,x} = \frac{\text{cov}(rk_a, rk_x)}{\sigma_{rk_a}\sigma_{rk_x}}, \tag{4.4}$$

where cov is the covariance and σ the standard deviation. An SRCC of ± 1 occurs if one sample is a perfect monotonic function of the other. Using Equation (4.4), we can then compute:

$$\widetilde{\rho}_\ell \coloneqq \max_{k \in \{1,\ldots,K\}} \{|\rho_{a_k,z_\ell}|\} \quad \forall \ell \in \{1,\ldots,L\} \tag{4.5}$$

as the maximum absolute rank correlation between a given legitimate feature z_ℓ and any protected feature a_k. We take the absolute values of SRCC in order to have $\widetilde{\rho}_\ell \in [0,1]$. However, our idea would also be consistent with, for instance, squaring the SRCC values instead. Our intuition behind taking the maximum over, for instance, the sum is that we do *not* want to penalize having many low individual absolute correlations—but rather scenarios where a seemingly legitimate feature is a (potentially noisy) proxy for one of the protected features.

Note that SRCC works for both numerical and ordinal features—this is important for our work. If, however, non-binary categorical features are present in $\mathcal{D}$, then traditional correlation measures are generally not a meaningful way of determining relationships. Alternatively, for instance, we might want to refer to the correlation ratio (Pearson, 1911), usually denoted by $\eta \in [0,1]$, which measures the relationship between *inter*-category variability and *intra*-category variability of some feature. As an example, assume we have a three-dimensional feature *gender* $\in \{F, M, O\}$ and want to quantify the relationship between gender and GRE V as well as GRE Q. Further assume that we have three observations per category (i.e., gender) and that the feature values are given as in Table 4.2 on page 91. By construction, the overall variability of GRE V scores is solely due to inter-category variability $(\eta = 1)$, whereas for GRE Q the category means are the same, hence $\eta = 0$. Because of the same value range and corresponding interpretation of η and $|\rho_{a_k,z_\ell}|$, we could straightforwardly adapt Equation (4.5) by replacing the latter with the former.

For simplicity, and because most traditional classification algorithms require encoding of categorical features as well, we assume in the following that $\mathcal{D}$ does *not* contain non-binary categorical features (e.g., because any such feature has been encoded

Tab. 4.2.: Exemplary data for illustrating correlation ratio.

ID	Gender	GRE V	GRE Q
1	*F*	130	140
2	*F*	130	150
3	*F*	130	160
4	*M*	150	140
5	*M*	150	150
6	*M*	150	160
7	*O*	170	140
8	*O*	170	150
9	*O*	170	160

accordingly) and that SRCC is applicable. We can then use $\widetilde{\rho}_\ell$ to further adjust the distance measure from Equation (4.3):

$$d''(i,\star) \coloneqq \sum_{\ell=1}^{L} \omega_\ell \left(1-\widetilde{\rho}_\ell\right)\left(1-z_\ell^{(i)}\right), \tag{4.6}$$

where high values of $\widetilde{\rho}_\ell$ reduce the importance of z_ℓ on the distance d''. Note that in the extreme case of $\widetilde{\rho}_\ell = 1$, the distance d'' will be independent of feature z_ℓ. This is desirable as it renders ineffective the possibility of introducing proxies for protected features under seemingly innocuous names.

For our graduate school admission example, we calculate the SRCC values using the `spearmanr` function of `SciPy` (Virtanen et al., 2020). Note that we only have one protected feature: *gender*. The absolute correlations are $\widetilde{\rho}_V = 0.035$, $\widetilde{\rho}_Q = 0.262$, and $\widetilde{\rho}_{AW} = 0.167$. While these values are not strikingly high, we may infer that there is a stronger relationship between *gender* and GRE Q than with GRE V or GRE AW. The importance of GRE Q for admission is thus reduced by 26.2%, as opposed to 3.5% and 16.7% for GRE V and GRE AW, respectively. In general, even if a seemingly legitimate feature was highly important for past decisions, its importance will vanish if it is highly correlated with a protected feature, as desired.

Coming back to Definition 4.1, we now define what *being more qualified* could mean:

Definition 4.3 (Higher qualification)**.** *We call an observation* (i) more qualified than (j) *if, according to the* $(\uparrow)$ *or* $(\downarrow)$ *relationships between features and positive outcome,* (i) *is better or equal than* (j) *for all legitimate features and strictly better for at least one* $\ell' \in \{1,\ldots,L\}$*, with* $\omega_{\ell'} \neq 0$ *and* $\widetilde{\rho}_{\ell'} \neq 1$.

Note that *being more qualified* is a stronger requirement than observation (i) having a shorter distance to the North Star than (j), that is, *being more qualified* implies a shorter distance to the North Star. The converse is not generally true. This implication is formally stated in the following proposition:

Proposition 4.1. *If, according to Definition 4.3, an observation* (i) *is more qualified than observation* (j)*, then* $d''(i, \star)$ *is strictly smaller than* $d''(j, \star)$*, where* d'' *is defined as in Equation* (4.6).

Proof. Assume (i) is more qualified than (j), and without loss of generality assume that all legitimate features are scaled as in Algorithm 1 on page 87. Then we have:

$$z_\ell^{(i)} \geq z_\ell^{(j)} \; \forall \ell \in \{1, \ldots, L\} \quad \text{and} \quad \exists \ell' \in \{1, \ldots, L\} : z_{\ell'}^{(i)} > z_{\ell'}^{(j)}.$$

With $\psi_\ell \coloneqq \omega_\ell \left(1 - \widetilde{\rho}_\ell\right) \in [0, 1]$ and $\psi_{\ell'} \neq 0$, we then obtain:

$$\begin{aligned} d''(i, \star) &= \sum_{\ell=1}^{L} \psi_\ell \left(1 - z_\ell^{(i)}\right) \\ &= \sum_{\ell=1}^{L} \psi_\ell - \left(\psi_1 z_1^{(i)} + \cdots + \psi_{\ell'} z_{\ell'}^{(i)} + \cdots + \psi_L z_L^{(i)}\right) \\ &< \sum_{\ell=1}^{L} \psi_\ell - \left(\psi_1 z_1^{(j)} + \cdots + \psi_{\ell'} z_{\ell'}^{(j)} + \cdots + \psi_L z_L^{(j)}\right) \\ &= d''(j, \star), \end{aligned}$$

since $\psi_{\ell'} z_{\ell'}^{(i)} > \psi_{\ell'} z_{\ell'}^{(j)}$ and $\psi_\ell z_\ell^{(i)} \geq \psi_\ell z_\ell^{(j)}$ for all other $\ell \in \{1, \ldots, L\} \setminus \{\ell'\}$. □

Note that in Table 4.1 on page 87, according to Definitions 4.1 and 4.3, observation 11 is treated unfairly over both observations 1 and 2. We will show that this can not happen with our method.

4.3.4 A Fair Ranking-Based Classification Algorithm

In this section, we summarize the previous findings and formalize our idea of a fair ranking-based classification algorithm. The proposed method can: a) fairly rank a given set of observations, b) propose new labels for the given observations, and c) rank and classify previously unseen observations. For a), we compute d'' for all observations and rank them by distance. After ranking, we reset the indices such

that observation (1) has the smallest distance to the North Star and (N) the largest. For b) and c), we need to define a capacity threshold $\alpha \in (0, 1)$. This could be, for instance, a given admission rate. Alternatively, we can set α to the share of positive outcomes within $\mathcal{D}$. Knowing α, we can then determine the cutoff point $\nu \coloneqq \lceil \alpha N \rceil$, such that the top-$\nu$ observations are assigned the positive outcome $(+)$ and the rest is assigned the negative outcome $(-)$. To infer a predictor, we compute

$$\delta \coloneqq \frac{(d''(\nu, \star) + d''(\nu + 1, \star))}{2} \tag{4.7}$$

as the average distance of observations (ν) and $(\nu + 1)$ to the North Star. Note that (ν) is the last observation with positive outcome, and $(\nu + 1)$ is the first observation with negative outcome.

Ultimately, to classify a previously unseen observation (u), we need to scale its legitimate features according to Algorithm 1 on page 87, using the minimum and maximum feature values as observed in $\mathcal{D}$—and measure the distance $d''(u, \star)$ to the North Star. The inferred predictor would then be:

$$\widehat{Y}^{(u)} = \begin{cases} (+) & \textit{if } \ d''(u, \star) \leq \delta \\ (-) & \textit{otherwise}. \end{cases} \tag{4.8}$$

The proposed method is summarized in Algorithm 2 on page 94. Note that from Proposition 4.1, it follows that meritocratic unfairness can not occur with our method:

Corollary 4.1. *Meritocratic unfairness, as stated in Definition 4.1, can* not *occur if observations are ranked and classified as in Algorithm 2 on page 94.*

Proof. From Proposition 4.1, we conclude that if (i) is more qualified than (j), then $d''(i, \star)$ will be strictly smaller than $d''(j, \star)$. But by construction of the ranking in Algorithm 2 on page 94, we will then have (i) ranked higher than (j), which also implies that if (j) is assigned $(+)$, then (i) as well. □

4.3.5 On the Relationship to Fairness Through Awareness

As explained in Section 4.2, *fairness through awareness* (Dwork et al., 2012) is one of the most prominent concepts of individual fairness, which is often verbalized as "treating similar individuals similarly." However, it is often not immediately clear

Algorithm 2: Fair ranking-based classification algorithm.

Input : Dataset $\mathcal{D}$; $(\uparrow)$ or $(\downarrow)$ relationships for legitimate features; threshold α.
Output : Ranked and classified observations $(1), \ldots, (N)$; predictor $\widehat{Y}$.

Compute Z as in Algorithm 1 on page 87;
Set $z_\ell^{(\star)} \leftarrow 1 \quad \forall \ell \in \{1, \ldots, L\}$;
Obtain $\omega_1, \ldots, \omega_L$ from learned classifier;
for $\ell \in \{1, \ldots, L\}$ **do**
  $\widetilde{\rho}_\ell \leftarrow \max_{k \in \{1,\ldots,K\}} \{|\rho_{a_k, z_\ell}|\}$ as in Equation (4.5);
end
for $i \in \{1, \ldots, N\}$ **do**
  $d''(i, \star) \leftarrow \sum_{\ell=1}^{L} \omega_\ell \left(1 - \widetilde{\rho}_\ell\right) \left(1 - z_\ell^{(i)}\right)$ as in Equation (4.6);
  Assign observation (i) the distance $d''(i, \star)$;
end
Rank observations by distance d'' and reset indices such that (1) has smallest distance;
Define cutoff point $\nu \leftarrow \lceil \alpha N \rceil$;
Assign $(+)$ to top-ν observations and $(-)$ to rest;
Define $\delta \leftarrow \frac{(d''(\nu, \star) + d''(\nu+1, \star))}{2}$;
if $d''(u, \star) \leq \delta$ *for a (potentially unseen) scaled observation* (u) **then**
  $\widehat{Y}^{(u)} \leftarrow (+)$;
else
  $\widehat{Y}^{(u)} \leftarrow (-)$;
end

Tab. 4.3.: Two observations with equal distance to the North Star.

	GRE V	GRE Q	GRE AW
Observation (i)	170	160	3.0
Observation (j)	140	160	6.0

how to measure *similarity* of individuals. Algorithm 2 on page 94 ranks observations based on their (weighted) distance to the North Star, $d''(\cdot, \star)$. Hence, by construction, if observations (i) and (j) have (relatively) similar distances to the the North Star, then their rankings will be similar as well. Specifically, for observations (i), (j), (k), and $rk_{(i)} > rk_{(j)} > rk_{(k)}$, with rk denoting the ranking of an observation, the following two inequalities will always hold:

$$d''(k, \star) - d''(i, \star) > d''(j, \star) - d''(i, \star) \tag{4.9}$$

$$d''(k, \star) - d''(i, \star) > d''(k, \star) - d''(j, \star). \tag{4.10}$$

However, having a similar distance to the North Star—hence, a similar ranking—does *not* imply that the respective observations are similar in a literal sense. For instance, in our running graduate school admission example, the two observations in Table 4.3 would have the same distance to the North Star, despite being fundamentally different in their feature values of GRE V and GRE AW. We argue that this is a desirable property, as it allows individuals with heterogeneous (but equally important and desirable) skill sets to achieve the positive outcome—at least to the extent that ω_ℓ and $\widetilde{\rho}_\ell$, $\ell \in \{1, \ldots, L\}$, allow.

On the other hand, it would be difficult to justify an algorithm that assigns significantly different outcomes to similar individuals. This is, in fact, the reasoning behind FTA. We will now show that our proposed method respects this requirement—precisely, that similar individuals are guaranteed to have similar distances to the North Star, and thus, similar rankings. But first, we define the similarity of two observations in terms of their weighted distance to each other in the feature space.

Definition 4.4 (Similarity of two observations)**.** *We measure the similarity of two observations* (i) *and* (j) *by their (weighted) taxicab distance to each other, similar to Equation* (4.6)*:*

$$d''(i, j) \coloneqq \sum_{\ell=1}^{L} \omega_\ell \left(1 - \widetilde{\rho}_\ell\right) \cdot \left| z_\ell^{(i)} - z_\ell^{(j)} \right|. \tag{4.11}$$

Again, we have the symmetry $d''(i,j) = d''(j,i)$. The following proposition now says that if two observations (i) and (j) are ε-similar, then the difference in their respective distances to the North Star will be bounded by ε.

Proposition 4.2. *If two observations (i) and (j) are ε-similar, that is, $d''(i,j) = \varepsilon$, $\varepsilon \geq 0$, then the following holds:*

$$\left|d''(i,\star) - d''(j,\star)\right| \leq \varepsilon.$$

Proof. Let $d''(i,j) = \varepsilon$, and define $\psi_\ell \coloneqq \omega_\ell \left(1 - \widetilde{\rho}_\ell\right)$. Then we have:

$$d''(i,j) = \sum_{\ell=1}^{L} \psi_\ell \cdot \left| z_\ell^{(i)} - z_\ell^{(j)} \right| = \varepsilon.$$

And further, with $\psi_\ell \geq 0$ and the triangle inequality:

$$\begin{aligned}
\left|d''(i,\star) - d''(j,\star)\right| &= \left| \sum_{\ell=1}^{L} \psi_\ell \left(1 - z_\ell^{(i)}\right) - \sum_{\ell=1}^{L} \psi_\ell \left(1 - z_\ell^{(j)}\right) \right| \\
&= \left| \sum_{\ell=1}^{L} \psi_\ell z_\ell^{(i)} - \sum_{\ell=1}^{L} \psi_\ell z_\ell^{(j)} \right| \\
&= \left| \sum_{\ell=1}^{L} \psi_\ell \left(z_\ell^{(i)} - z_\ell^{(j)} \right) \right| \\
&\leq \sum_{\ell=1}^{L} \psi_\ell \cdot \left| z_\ell^{(i)} - z_\ell^{(j)} \right| = \varepsilon.
\end{aligned}$$

This shows the desired result. □

Now, if we let ε become small, that is, $\varepsilon \to 0$, then the observations *and* their respective distances to the North Star are becoming increasingly similar—and in the limit equal. Hence, those observations will be ranked adjacently, everything else unchanged.

4.4 Case Study: German Credit Dataset

In this section, we instantiate our proposed method on the widely-used German Credit dataset (Dua & Graff, 2017). The dataset is made up of 1000 observations classified as *good* (70%) or *bad* (30%) credits (Y). As summarized by Pedreshi

et al. (2008), it includes 20 features on a) personal belongings (e.g., *checking account status*, *savings status*, *property*), b) past/current credits and requested credit (e.g., *credit history*, *credit request amount*), c) employment status (e.g., *job type*, *employment since*), and d) personal attributes (e.g., *personal status and gender*, *age*, *foreign worker*).

Tab. 4.4.: Features of the German Credit dataset after pre-processing.

Feature	**Description**	A **or** X	$(\uparrow)$ **or** $(\downarrow)$	ω_ℓ	$\widetilde{\rho}_\ell$
personal status and gender	marital status and gender	A	–	–	–
age	age of person	A	–	–	–
foreign worker	foreign worker yes/no	A	–	–	–
checking account status	money in checking	X	$(\uparrow)$	$0.28\ (\sigma = 0.047)$	0.07
savings status	money in savings	X	$(\uparrow)$	$0.11\ (\sigma = 0.029)$	0.04
property	value of property	X	$(\uparrow)$	$0.13\ (\sigma = 0.039)$	0.13
type of housing	free/rent/own	X	$(\uparrow)$	$0.06\ (\sigma = 0.025)$	0.07
credit history	quality of credit history	X	$(\uparrow)$	$0.11\ (\sigma = 0.032)$	0.15
credit request amount	credit amount requested	X	$(\downarrow)$	$0.18\ (\sigma = 0.042)$	0.05
job type	unempl./un-/skilled/mgmt.	X	$(\uparrow)$	$0.04\ (\sigma = 0.018)$	0.11
employment since	how long employed	X	$(\uparrow)$	$0.09\ (\sigma = 0.031)$	0.32

From the original dataset, we exclude certain features—such as *telephone*—from consideration as they do not exhibit an obvious monotonic relationship with the outcome and, more importantly, appear to be irrelevant for deciding whether to grant a loan or not. The remaining features are shown in Table 4.4. Similar to existing literature, we further separate the remaining features into protected and legitimate features. We determine the relationships ($\uparrow$ or $\downarrow$) as depicted in Table 4.4. For evaluation purposes later on, we randomly shuffle the data and set aside 200 observations for testing purposes, 150 of which are labeled as having *good* credit.

Experimental Setup First, we scale the legitimate features X as in Algorithm 1 on page 87. Then, we fit a random forest classifier with bootstrapping to predict Y from X. We repeat this five times, and for each model, we randomly permute the features ten times—this results in 50 estimates of importance for each legitimate feature. The average numbers (including standard deviations) are displayed as ω_ℓ in Table 4.4. Note that the values of feature importance are only meaningful if the underlying model predicts Y reasonably well. In our case, we obtain average accuracies of 79.4% (training) and 78.7% (testing). Following Algorithm 2 on page 94, we next compute the maximum absolute rank correlations $\widetilde{\rho}_\ell$ for each legitimate feature (see Table 4.4).

We conduct several experiments to rank 200 test observations and predict *good* or *bad* credit. To that end, we train a logistic regression classifier for the following scenarios: a) using all available features, including protected features (`LogReg all`),

Tab. 4.5.: Meritocratic unfairness and accuracy of different scenarios for the German Credit dataset.

	`LogReg all`	`LogReg FTU`	Our method	Test labels
$\mathcal{S}$	57.5% $\left(\frac{115}{200}\right)$	56.0% $\left(\frac{112}{200}\right)$	0.0%	14.5% $\left(\frac{29}{200}\right)$
$\mathcal{T}$	616	596	0	222
Accuracy	78.5%	76.5%	56.0%	100%

and b) omitting protected features (`LogReg FTU`). Third, c) we apply our proposed method to the test observations.

Evaluation Criteria and Results To evaluate the results from scenarios a)–c), we first compare the rankings induced by the respective methods: For a) and b), we rank observations based on the prediction probabilities returned by the classifier, and for c), the ranking is obtained as in Algorithm 2 on page 94. We measure fairness of a ranking by the number of unfairly treated observations, as specified in Definitions 4.1 and 4.3. In general, an observation can be treated unfairly over *more* than one other observation.

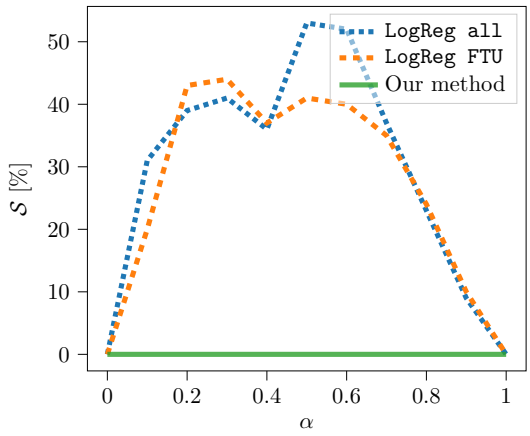
For the baseline models, we therefore measure both the share $\mathcal{S}$ of individual observations that are treated unfairly and the total number $\mathcal{T}$ of instances where meritocratic unfairness occurs. The results are depicted in Table 4.5. Additionally, we also provide numbers on the meritocratic unfairness of the test labels—where we assume that observations with *good* credit are ranked higher than observations with *bad* credit. We note that `LogReg all` produces both the highest $\mathcal{S}$ and the highest $\mathcal{T}$, and `LogReg FTU` performs only marginally better.

In reality, observations will primarily be affected by the actual outcome of the decision-making task—*good* or *bad* credit. Hence, we also compare scenarios a)–c) with respect to the predicted outcome, dependent on the choice of a threshold α. Specifically, we calculate $\mathcal{S}$ for $\alpha \in \{0, 0.1, 0.2, \ldots, 1\}$ and each scenario. Note that for the logistic regression models, the positive outcome (*good* credit) is assigned to the $(100 \cdot \alpha)\%$ observations with the highest prediction probabilities. From Figure 4.1 on page 99, we conclude that, apart from the trivial cases of $\alpha = 0$ and $\alpha = 1$, both baseline models involve high percentages of unfairly treated observations—with `LogReg all` reaching values of more than 50% for $\alpha = 0.5$ and $\alpha = 0.6$.

For completeness, we also include the models' accuracy with respect to the test labels in Table 4.5.[2] However, note that accuracy is measured based on an imperfect and

[2]We set $\alpha = 0.75$ to ensure comparability with the test labels.

Fig. 4.1.: Share $\mathcal{S}$ of unfairly treated observations over α for different scenarios.



Note: $\mathcal{S}$ is calculated based on the predicted labels*, not the ranking.*

potentially biased proxy (i.e., the test labels) of the ground-truth labels regarding the qualification of individuals. Hence, a drop in accuracy, as observed for our method in Table 4.5 on page 98, may be explained by a strong mismatch between available imperfect labels and true (but unavailable) labels. This trade-off between accuracy and fairness is often referred to as the "cost of fairness" (Corbett-Davies et al., 2017; von Zahn et al., 2022). Unfortunately, we do not have a way to control the level of label bias in real-world data. For that reason, we conduct a series of experiments on synthetic data and present evidence that, in fact, our method's accuracy tends to be a) similar to traditional classification models when label bias is low, and b) lower when label bias is high, implying that low accuracy and desirable outcomes need not always contradict each other.

4.5 Experiments on Synthetic Data

In order to better understand the previous results, we evaluate our method extensively on synthetic data with imperfect labels. To that end, we take a more in-depth look at the simplified graduate school admission data introduced in Section 4.3. Recall that we sampled the GRE scores from multivariate Gaussian distributions according to the gender-specific means and standard deviations provided by ETS (2019, 2018). Also, we included an equal amount of women and men, respectively, in the dataset of overall 1000 observations.

Experimental Setup For the purpose of evaluating our method, we simulate historical admission decisions/labels (e.g., of a technical university) first by computing a score R for each observation as the weighted sum of its (scaled) feature values:

$$R := \frac{\zeta}{\zeta+4} \cdot \mathbb{1}_{male} + \frac{1}{\zeta+4} \cdot \text{GRE V} + \frac{2}{\zeta+4} \cdot \text{GRE Q} + \frac{1}{\zeta+4} \cdot \text{GRE AW} + \epsilon,$$

with $\zeta \geq 0$ and noise $\epsilon \in \mathcal{N}(0, 0.1)$, the latter of which might reflect the (unpredictable) mood of the admissions committee or other circumstances that affected admission decisions in the past. Note that the feature weights sum up to 1, and that the weights of GRE V and GRE AW are the same. Moreover, the influence of GRE Q on R is approximately twice as high as compared to the other GRE scores, in order to mimic a more quantitative-focused admission process. The positive outcome $(+)$ is then initially assigned to observations with $R > 0.5$, and $(-)$ is assigned otherwise—this ensures a well-balanced label distribution. Yet, those generated labels are imperfect (i.e., *not* ground truth) because a) the score R is only a noisy signal of potential success in graduate school, b) the computation of R involves (simulated) human subjectivity and error, and c) R may be discriminatory, depending on the choice of ζ.

The parameter ζ lets us control the amount of *direct* discrimination (Mehrabi et al., 2021) in the decisions, as it directly increases R for males and decreases it for females. Besides, a large ζ could also be an indicator of *indirect* discrimination (Mehrabi et al., 2021), for instance, if other features highly correlated with gender—and favoring males—were given strong weight in the (simulated) historical decisions. Note that as ζ becomes increasingly large, the direct influence of the legitimate features (GRE scores) on R vanishes.

We evaluate our method on seven synthetic datasets with varying levels of bias/discrimination in the labels, as controlled through ζ (see Table 4.6 on page 101). Like in Section 4.4, we randomly set aside 200 observations for testing on each dataset. Our method is implemented according to Algorithm 2 on page 94, with the GRE features being legitimate, and gender being protected. Naturally, higher GRE scores should be more beneficial towards being admitted—hence $(\uparrow)$ relationships with the outcome. An overview of all ω_ℓ and $\widetilde{\rho}_\ell$ values is given in Table 4.7 on page 102. Note that $\widetilde{\rho}_\ell$ is constant across the datasets, as changing ζ only affects the label distribution, not the correlations among features. We also like to highlight that feature importances (ω_ℓ) still capture well the policy that GRE Q should carry significantly more weight in the decision process than GRE V and GRE AW—even with relatively high levels of bias in the labels $(\zeta = 3.0)$.

Tab. 4.6.: Meritocratic unfairness, accuracy, and admission statistics on synthetic data with varying levels of discrimination ζ.

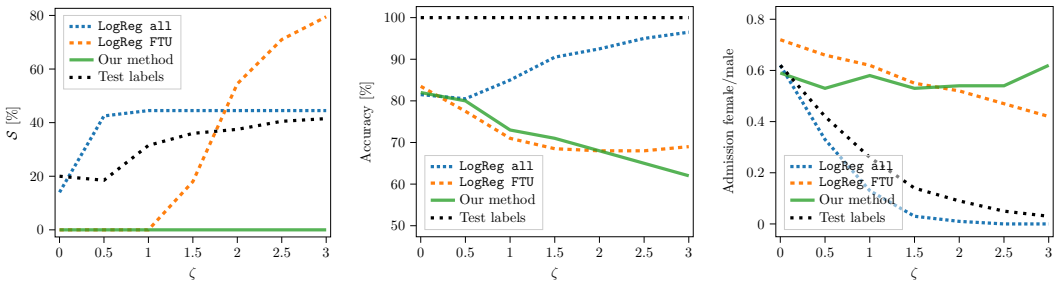
		`LogReg all`	`LogReg FTU`	Our method	Test labels
	$\mathcal{S}$	14.0% $\left(\frac{28}{200}\right)$	0.0%	0.0%	20.0% $\left(\frac{40}{200}\right)$
	$\mathcal{T}$	45	0	0	177
$\zeta = 0.0$	Accuracy	81.5%	83.5%	82.0% $(\alpha = 0.59)$	100.0%
	Admission female/male	0.62	0.72	0.59 $(\alpha = 0.59)$	0.62
	Admission rate female	55.4%	59.8%	47.8% $(\alpha = 0.59)$	48.9%
	Admission rate male	75.9%	70.4%	68.5 $(\alpha = 0.59)$	67.6%
	$\mathcal{S}$	42.5% $\left(\frac{85}{200}\right)$	0.0%	0.0%	18.5% $\left(\frac{37}{200}\right)$
	$\mathcal{T}$	503	0	0	230
$\zeta = 0.5$	Accuracy	80.5%	77.5%	80.0% $(\alpha = 0.56)$	100.0%
	Admission female/male	0.33	0.66	0.53 $(\alpha = 0.56)$	0.42
	Admission rate female	32.6%	53.3%	42.4% $(\alpha = 0.56)$	35.9%
	Admission rate male	84.3%	68.5%	67.6% $(\alpha = 0.56)$	73.1%
	$\mathcal{S}$	44.5% $\left(\frac{89}{200}\right)$	0.0%	0.0%	31.5% $\left(\frac{63}{200}\right)$
	$\mathcal{T}$	$1,200$	0	0	608
$\zeta = 1$	Accuracy	85.0%	71.0%	73% $(\alpha = 0.60)$	100.0%
	Admission female/male	0.13	0.62	0.58 $(\alpha = 0.60)$	0.26
	Admission rate female	14.1%	48.9%	47.8% $(\alpha = 0.60)$	27.2%
	Admission rate male	91.7%	67.6%	70.4% $(\alpha = 0.60)$	88.0%
	$\mathcal{S}$	44.5% $\left(\frac{89}{200}\right)$	18.0% $\left(\frac{36}{200}\right)$	0.0%	36.0% $\left(\frac{72}{200}\right)$
	$\mathcal{T}$	$1,571$	48	0	797
$\zeta = 1.5$	Accuracy	90.5%	68.5%	71.0% $(\alpha = 0.56)$	100.0%
	Admission female/male	0.03	0.55	0.53 $(\alpha = 0.56)$	0.14
	Admission rate female	3.3%	41.3%	42.4% $(\alpha = 0.56)$	15.2%
	Admission rate male	96.3%	63.9%	67.6% $(\alpha = 0.56)$	90.7%
	$\mathcal{S}$	44.5% $\left(\frac{89}{200}\right)$	54.5% $\left(\frac{109}{200}\right)$	0.0%	37.5% $\left(\frac{75}{200}\right)$
	$\mathcal{T}$	$1,630$	324	0	955
$\zeta = 2$	Accuracy	92.5%	68.0%	68.0% $(\alpha = 0.55)$	100.0%
	Admission female/male	0.01	0.52	0.54 $(\alpha = 0.55)$	0.09
	Admission rate female	1.1%	41.3%	41.3% $(\alpha = 0.55)$	9.8%
	Admission rate male	99.1%	67.6%	65.7% $(\alpha = 0.55)$	92.6%
	$\mathcal{S}$	44.5% $\left(\frac{89}{200}\right)$	71.0% $\left(\frac{142}{200}\right)$	0.0%	40.5% $\left(\frac{81}{200}\right)$
	$\mathcal{T}$	$1,631$	731	0	$1,221$
$\zeta = 2.5$	Accuracy	95.0%	68.0%	65.0% $(\alpha = 0.54)$	100.0%
	Admission female/male	0.00	0.47	0.54 $(\alpha = 0.54)$	0.05
	Admission rate female	0.0%	37.0%	41.3% $(\alpha = 0.54)$	5.4%
	Admission rate male	100.0%	66.7%	64.8% $(\alpha = 0.54)$	95.4%
	$\mathcal{S}$	44.5% $\left(\frac{89}{200}\right)$	79.5% $\left(\frac{159}{200}\right)$	0.0%	41.5% $\left(\frac{83}{200}\right)$
	$\mathcal{T}$	$1,631$	$1,364$	0	$1,257$
$\zeta = 3$	Accuracy	96.5%	69.0%	62.0% $(\alpha = 0.54)$	100.0%
	Admission female/male	0.00	0.42	0.62 $(\alpha = 0.54)$	0.03
	Admission rate female	0.0%	32.6%	44.6% $(\alpha = 0.54)$	3.3%
	Admission rate male	100.0%	65.7%	61.1% $(\alpha = 0.54)$	96.3%

Tab. 4.7.: Overview of seven synthetic datasets with varying levels of discrimination ζ.

			$\zeta = 0.0$	$\zeta = 0.5$	$\zeta = 1.0$	$\zeta = 1.5$	$\zeta = 2.0$	$\zeta = 2.5$	$\zeta = 3.0$	
Feature	A or X	($\uparrow$) or ($\downarrow$)	ω_ℓ	ω_ℓ	ω_ℓ	ω_ℓ	ω_ℓ	ω_ℓ	ω_ℓ	$\widetilde{\rho}_\ell$
Gender	A	–	–	–	–	–	–	–	–	–
GRE V	X	($\uparrow$)	0.26	0.21	0.22	0.21	0.22	0.24	0.27 ($\sigma = 0.04$)	0.04
GRE Q	X	($\uparrow$)	0.66	0.71	0.72	0.70	0.63	0.56	0.49 ($\sigma = 0.05$)	0.24
GRE AW	X	($\uparrow$)	0.08	0.08	0.07	0.09	0.16	0.20	0.24 ($\sigma = 0.03$)	0.13

Results and Interpretation As previously, we first compare the rankings of our method against the rankings induced by the logistic regression models `LogReg all` and `LogReg FTU` on the test data. The resulting levels of meritocratic unfairness (both $\mathcal{S}$ and $\mathcal{T}$) are displayed in Table 4.6 on page 101, including the statistics for the test labels. We also report accuracy as well as additional statistics regarding the admission of women and men for each scenario, based on label predictions. Specifically, we report the ratio of admitted women to men and compare the admission rates per gender. For label predictions with our method, to ensure comparability, we set α equal to the share of admitted applicants in the test set.

Fig. 4.2.: Empirical results of experiments on synthetic data.



(a) (b) (c)

Note: Part (a) displays the share $\mathcal{S}$ of unfairly treated observations over the degree of discrimination ζ. Part (b) displays accuracy over ζ. Part (c) displays the admission ratio of females to males over ζ.

From Table 4.6 on page 101 and Figure 4.2, we infer several observations: first, as ζ increases, the admission ratio of females to males in the datasets decreases, as expected, whereas the overall admission rate remains stable (between 54% and 60% in the test labels). The fact that increasing ζ results in the labels depending stronger on the value of *gender* is exploited by `LogReg all` to discriminate observations based thereon. Not surprisingly, as ζ increases, `LogReg all` clings to the trajectory of the test labels both for accuracy as well as meritocratic unfairness and demographic parity (with respect to admission ratios), making its predictions accurate but blatantly unfair—both meritocratically and with respect to admission

rates by gender. We further observe in Figure 4.2 (a) on page 102 that the gender-agnostic `LogReg FTU` model removes meritocratic unfairness when label bias is low ($\zeta < 1.0$). However, as the level of discrimination in the data increases, it fails to remove such unfairness, with $\mathcal{S}$ surging and even surpassing the levels of `LogReg all` and the test labels for growing ζ. These problems do not occur with our method, which always achieves zero meritocratic unfairness. Additionally, as can be seen in Figure 4.2 (c) on page 102, our experiments suggest that enforcing individual meritocratic fairness results in higher group fairness (here: demographic parity with respect to admission rates) as well: while the logistic regression models both exhibit a negative relationship between ζ and demographic parity, our method satisfies a constant high level of group fairness for any ζ, similar to the one of the test labels without explicit discrimination (0.62 for $\zeta = 0.0$). Note that the converse—group fairness implying individual meritocratic fairness—is not generally true, for instance, if a model randomly admits an equal share of women and men without paying any attention to their qualification.

4.6 Conclusion

In this work, we presented a practical and easy-to-implement approach for fair ranking and binary classification based on monotonic relationships between legitimate features (or interactions thereof) and the outcome. Given the common setting of data with (potentially biased) imperfect labels, our method ranks observations according to their qualification for a specific outcome, for instance, admission to graduate school, regardless of protected features like gender or race. Instead of learning to predict imperfect labels, we introduce an idea to incorporate useful information from historical decisions in our decision criterion. Additionally, we account for unwanted dependencies between (seemingly) legitimate and protected features. We show theoretically that our method respects a version of the prominent concept of *fairness through awareness*, as proposed by Dwork et al. (2012). Experiments on synthetic and real-world data confirm that our method yields desirable results both with respect to meritocratic fairness and group fairness (e.g., similar admission rates for females and males), clearly outperforming traditional classification algorithms trained on data with biased labels.

Our work involves certain limitations that allow for various directions of follow-up research: for instance, it would be interesting to elaborate more on how to meaningfully include features that do not exhibit obvious monotonic relationships with the outcome. Another natural extension of our method could involve a more

sophisticated and natural way of accounting for feature interactions. Perhaps most importantly, we stress that our proposed method satisfies the introduced conception of meritocratic fairness as well as demographic parity (empirically), but may *not* be fair according to other notions of fairness. More generally, we like to highlight that fairness in the societal sense cannot be reduced to a simple technical property. For instance, understanding people's fairness *perceptions* towards algorithmic decision systems is another vital research endeavor (M. K. Lee, 2018; Schöffer et al., 2021). Hence, it is necessary to have different notions of fairness—and our work may be seen as one contribution to this toolbox. Ultimately, we hope that our work will equip (especially) practitioners with helpful new tools for designing responsible AI systems.

5 Assessing Fairness Perceptions Towards AI Systems

In this chapter, we carry out a mixed-method online study to explore people's perceptions of informational fairness and trustworthiness towards artificial intelligence (AI) systems when presented with varying degrees of system-related information. Specifically, we instantiate an AI system for automated loan approval and generate diverse explanations that are frequently utilized in academic literature. We then randomize the amount of information that study participants are exposed to. Our quantitative analyses reveal that both the quantity of information and individuals' self-assessed AI literacy significantly impact the perceived informational fairness, which in turn positively correlates with the perceived trustworthiness of the AI system. A thorough analysis of qualitative feedback illuminates people's expectations for explanations, which include (i) consistency, (ii) disclosure of monotonic relationships between features and outcome, and (iii) actionability of recommendations. Note that in this and the following chapter, we use the term *automated decision system* when referring to an AI system that makes fully automated decisions.

5.1 Introduction

AI-informed decision-making has become ubiquitous in many high-stakes domains such as hiring (Kuncel et al., 2014), bank lending (Townson, 2020), grading (Satari-

This chapter is based on published work. To enhance the reading experience and maintain overall consistency of the thesis, we removed the abstract and made several minor adjustments. The original paper can be accessed via:

Schöffer, J., Kühl, N. & Machowski, Y. (2022). "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1616–1628). https://doi.org/10.1145/3531146.3533218.

ano, 2020), and policing (Heaven, 2020), among others. The underlying motives of adopting automated decision systems (ADSs) are manifold: they range from cost-cutting to improving performance and enabling more robust and objective decisions (Harris & Davenport, 2005; Kuncel et al., 2014; Newell & Marabelli, 2015). Hopes are also that, if properly designed, ADSs can be a valuable tool for breaking out of vicious patterns of human stereotyping and contributing to social equity, for instance, in the realms of recruitment (Chalfin et al., 2016; Koivunen et al., 2019), healthcare (Grote & Berens, 2020; Triberti et al., 2020), or financial inclusion (Lepri et al., 2017). However, ADSs are typically based on machine learning (ML) techniques, which, in turn, rely on historical data. If, for instance, this underlying data is biased (e.g., because certain socio-demographic groups were favored in a disproportionate way), an ADS will learn from and perpetuate existing patterns of unfairness (Feuerriegel et al., 2020). Prominent examples of such behavior from the recent past are race and gender stereotyping in job ad delivery (Imana et al., 2021), as well as the discrimination of Latinx and African-American borrowers in algorithmic mortgage loan pricing (Bartlett et al., 2022). These and other cases have put ADSs under enhanced scrutiny, justifiably jeopardizing trust in these systems (Edelman, 2021).

In recent years, a growing body of AI and ML research has been devoted to detecting, quantifying, and mitigating unfairness in ADSs (Mehrabi et al., 2021). A significant share of this work has focused on formalizing different concepts of *fairness* through statistical equity constraints, many of which are at odds with each other (Chouldechova, 2017; Kleinberg et al., 2017). As a consequence, there cannot be a one-size-fits-all technical fairness criterion. Moreover, in many cases, these techno-centric works do not explicitly take into account the opinions of people that are (potentially) affected by such automated decisions. While the Fairness, Accountability, and Transparency (FAccT) community has made a plethora of impactful contributions over the past years, it is still crucial to better understand people's perceptions and attitudes towards ADSs—in addition to how researchers may define those systems' fairness in technical terms.

A related issue revolves around explaining automated decisions to affected individuals. As ADSs employ ever more sophisticated and "black-box" ML models, several problems arise; one of which is the hampered detectability of adverse behavior of such systems. In order to safeguard transparency and accountability of automated decisions, several laws and regulations demand a "right to explanation" (Goodman & Flaxman, 2017). The General Data Protection Regulation (GDPR), for instance, requires the disclosure of "the existence of automated decision-making, including [...] meaningful information about the logic involved [...]" (European Union, 2016,

Section 2, Article 13) to data subjects. In fact, it has been shown, among others, that explanations can enhance people's understanding of certain automated decisions (Lim et al., 2009). For most real-world cases, however, those regulations generally remain (too) vague and little actionable, which often results in deficient adoption, as noticed in the context of bank lending (Szczygieł, 2022). Moreover, research on explainable AI (XAI) suggests that there exists no one-size-fits-all approach to explaining ADSs either (Arya et al., 2021; Langer, Oster, et al., 2021).

In this work, we conduct a human subject study to examine the effects of explanations on people's perceptions towards an automated loan approval system, where we randomize the type and amount of information that study participants get to see. The primary dependent variables that we are interested in are perceptions of *informational fairness* of the system (i.e., whether people think they are given adequate information on and explanation of the decision-making process and its outcomes) as well as perceived *trustworthiness*, and the relationship between both. We also assess the influence of people's (self-assessed) *AI literacy* on the outcomes. Finally, we ask multiple open-ended questions with respect to people's ability to assess the given system's fairness, as well as regarding the appropriateness of explanations' content.

5.2 Background

Topics of *fairness* and *trustworthiness* have become important pillars of AI and human-computer interaction (HCI) research in recent years. In this section, we provide an overview of relevant literature and highlight our contributions. It is—albeit unsurprisingly—important to note that a "fair" (according to some technical fairness notion) system does not imply that people perceive it as such; either because their personal fairness concepts differ from the employed technical notion or because they are not enabled to assess the system's (un)fairness to begin with. In fact, it must be questioned whether an ADS that satisfies given statistical notions of fairness (e.g., equitable distribution of outcomes) can ever be *truly* considered fair when at the same time decision subjects are left in the dark with respect to the inner workings of the system. Instead, *fairness* (of ADSs) is likely a multi-faceted construct that encompasses different dimensions, similar to dimensions of (organizational) justice, which are commonly made up of distributive, procedural, interpersonal, and informational justice (Colquitt & Rodell, 2015). While distributive and procedural aspects have been considered in the context of ADSs (e.g., by Grgić-Hlača, Zafar,

et al. (2018), M. K. Lee et al. (2019), and R. Long (2021)), work on *informational fairness* is lacking.

Borrowing from D. Chan (2011), we call a system *informationally fair* if it conveys adequate information on and explanation of the decision-making process and its outcomes; and we define *adequate information* (similar to Colquitt and Rodell (2015)) as information being thorough, reasonable, tailored to individual needs, as well as helping people understand the decision-making process, and enabling them to judge whether this process is fair or unfair. *Trustworthiness* is a well-established construct that, according to Bélanger et al. (2002, p. 252), is defined as "the perception of confidence in the [...] reliability and integrity [of an ADS]." We refer the reader to Jacovi et al. (2021), J. D. Lee and See (2004), and Vereschak et al. (2021) for survey literature on trust and trustworthiness.

5.2.1 Related Work

Automated Decision Systems Harris and Davenport (2005) define *automated decision systems* (ADSs) as systems that aim to minimize human involvement in decision-making processes. In this work, we assume ADSs to be supervised ML models. In many cases, ADSs have the potential to make more consistent decisions than humans. Such systems are popular in many industries, such as banking (Harris & Davenport, 2005; Townson, 2020) or hiring (Carey & Smith, 2016; Chalfin et al., 2016; Koivunen et al., 2019; Kuncel et al., 2014), and they are emerging in new areas as well, for instance, in healthcare (Grote & Berens, 2020; Triberti et al., 2020). With their increasing adoption in different consequential areas, it is important to ensure that ADSs reach fair decisions that are transparent, primarily, to affected individuals or auditors. However, there have been multiple cases in the recent past where algorithms made biased decisions that discriminated against certain groups, for instance, based on gender or race (Angwin et al., 2016; Buolamwini & Gebru, 2018; Heaven, 2020). In other instances, ADSs have been operating in an opaque ("black-box") fashion, making it, among others, difficult (i) for affected individuals to grasp the rationale behind certain decisions, and (ii) for regulatory agencies and other responsible stakeholders to vet such systems appropriately (Pasquale, 2015). On that account, fairness and transparency of ADSs have become important topics of interest for the research community.

Explainable AI (XAI) Despite being a popular topic of current research, XAI is a natural consequence of designing ADSs and, as such, has been around at least

since the 1980s (Lewis & Mack, 1982). Its importance, however, keeps rising as increasingly sophisticated (and opaque) AI techniques are used to inform ever more consequential decisions. Transparency is not only required by law (e.g., GDPR); Eslami et al. (2019), for instance, have shown that users' attitudes towards algorithms change when transparency is increased. In general, both quantity and quality of explanations matter: Kulesza et al. (2013) explored the effects of soundness and completeness of explanations on end users' mental models and suggest, among others, that oversimplification is problematic. Recent findings from Langer, Baum, et al. (2021), on the other hand, suggest that in the case of automated job interviews it might make sense to withhold certain pieces of information from applicants in order to not evoke negative reactions.

Even in the presence of explanations, people sometimes rely too heavily on system suggestions (Bussone et al., 2015), a phenomenon commonly referred to as *automation bias* (De-Arteaga et al., 2020; Goddard et al., 2014). Ehsan and Riedl (2021) have also used the term "explainability pitfalls" for any such unanticipated negative effects of explanations (e.g., unwarranted trust (Schlicker & Langer, 2021)). Eventually, Chromik et al. (2019) warn that explanations can be exploited to purposefully deceive users for the benefit of other stakeholders. Hence, explanations are by no means a silver bullet when it comes to solving problems of opaque AI systems (Bauer et al., 2023). A comprehensive overview of XAI stakeholders and their distinct desiderata is given by Langer, Oster, et al. (2021). For instance, people affected by automated decisions may be particularly interested in explanations that enable them to evaluate the fairness and trustworthiness of the underlying systems (Langer, Oster, et al., 2021; Schöffer & Kühl, 2021). This desideratum is closely linked to informational fairness of ADSs, as introduced earlier. We refer the interested reader to, among others, Adadi and Berrada (2018), Arrieta et al. (2020), Goebel et al. (2018), Guidotti et al. (2018), Langer, Oster, et al. (2021), T. Miller (2019), and Molnar (2020) for more in-depth literature on different XAI techniques and their inner workings. Regarding the effectiveness of explanations, generally speaking, prior research has primarily focused on comparing individual explanation styles head-to-head (e.g., Binns et al. (2018) and Dodge et al. (2019)), while little work has been done on evaluating the interplay of different styles, including potential complementarity. Langer, Oster, et al. (2021) emphasize the sparsity of empirical work with respect to the effectiveness of explanations overall.

Perceptions Towards ADSs A relatively new line of research in AI and HCI has started focusing on perceptions of fairness and trustworthiness in automated decision-making. For instance, Binns et al. (2018) and Dodge et al. (2019) compare fairness

perceptions towards ADSs for distinct explanation styles. Their works suggest differences in effectiveness of individual explanation styles—however, they also note that there does not seem to be a single best approach to explaining automated decisions. A different line of research has examined people's moral judgments with respect to the use of specific features in ADSs (Grgić-Hlača, Redmiles, et al., 2018; Grgić-Hlača, Zafar, et al., 2018), also with mixed empirical findings. M. K. Lee (2018) compares perceptions of fairness and trustworthiness depending on whether the decision maker is a person or an algorithm in the context of managerial decisions. Their findings suggest that, among others, people perceive automated decisions as less fair and trustworthy for tasks that require typical human skills. M. K. Lee and Baykal (2017) explore how algorithmic decisions are perceived in comparison to group-made decisions. R. Wang et al. (2020) combine a number of manipulations, such as favorable and unfavorable outcomes, to gain an overview of fairness perceptions. An interesting finding by M. K. Lee et al. (2019) suggests that fairness perceptions decline for some people when gaining an understanding of an algorithm if their personal fairness concepts differ from those of the algorithm. Woodruff et al. (2018) conducted workshops with people from traditionally marginalized backgrounds, inferring that awareness of unfairness in ADSs can substantially affect trust in companies or products.

Some work has also assessed the impact of people's demographics (including gender (Pierson, 2017)), as well as political views and task experience (Grgić-Hlača et al., 2022) on their perceptions. Saxena et al. (2020) examined lay people's perceptions of different technical fairness notions for ADSs, suggesting that people prefer notions related to meritocratic fairness (Joseph et al., 2016; Y. Liu et al., 2017). Regarding trustworthiness, Kizilcec (2016), for instance, concludes that it is important to provide the right amount of transparency for optimal trust effects, as both too much and too little transparency can have undesirable effects. Kästner et al. (2021) also examined the relationship between explainability and trust(worthiness), urging system designers to engineer for trustworthiness (as opposed to trust), and indicating that explanations can be a crucial toolbox towards that goal. Regarding perceptions of different social groups, M. K. Lee and Rich (2021) point out that prior studies have mostly recruited respondents from Amazon Mechanical Turk (Paolacci et al., 2010), which has predominantly White participants (Hitlin, 2016). Because of this, among other reasons (Prolific, 2022), we have recruited our study participants through Prolific (Palan & Schitter, 2018).

5.2.2 Research Gaps and Our Contributions

We aim to complement prior work to better understand how much of which information should be provided so that people are optimally enabled to understand the inner workings and appropriately assess the fairness and trustworthiness of ADSs. To that end, we conducted a randomized experiment to examine people's perceptions of informational fairness and trustworthiness towards an automated loan approval system, given different combinations of common explanations (relevant factors, factor importance, and counterfactual explanations). While there exists prior work on trustworthiness perceptions for *individual* explanation styles, we see a significant gap with respect to assessing *combinations* of different explanations. We argue that this is an important gap to fill because different explanations convey different information and will likely have to be leveraged complementarily (i.e., not in isolation) in practice. On a related note, we also set about examining the marginal effects of providing certain explanations *on top of* others—which, to the best of our knowledge, has not been analyzed in depth before. As a consequence, we alter the amount of information that different groups of people get to see. We do by no means claim to examine these aspects exhaustively, but we hope that our work will be a stepping stone for further research.

Finally, and perhaps most importantly, we shift focus from examining distributive and procedural fairness perceptions to *informational fairness*. In other words, we do *not* ask people whether they find particular ADS outcomes or procedures fair or not, but—broadly speaking—whether they feel they received sufficient information to *assess* a given system. This is an important distinction. Only very few works have considered the informational fairness dimension when experimentally evaluating effectiveness of ADS explanations: Binns et al. (2018) only measure the understandability aspect of informational fairness for individual explanation styles; Schlicker et al. (2021) assess informational fairness perceptions, but with a focus on comparing human with automated decision makers. Uhde et al. (2020) and A. Brown et al. (2019) conducted interviews and workshops to infer qualitative statements related to informational fairness; whereby A. Brown et al. (2019, p. 10) explicitly state that "more research is needed to understand how different elements of algorithmic systems affect perceptions of [...] informational justice." Empirical work on the interplay of informational fairness and trustworthiness perceptions for ADSs is, to our knowledge, entirely novel. Finally, we also analyze the relationship between study participants' (self-assessed) AI literacy and their perceptions, and we qualitatively examine their answers to open-ended question regarding (in)appropriateness of

explanations as well as what information they feel is missing (if any) to properly vet the given ADS.

5.3 Research Hypotheses

The conditions of our experiment comprise different amounts of information that study participants get to see with respect to an ADS in the realm of automated loan decisioning. Regarding the potential effects of varying amounts of information on our dependent variables of perceived informational fairness and trustworthiness, we formulate two research hypotheses based on preliminary qualitative insights with respect to people's desire for transparency and information (A. Brown et al., 2019; Uhde et al., 2020) as well as prior findings from the psychology literature (Colquitt & Rodell, 2011, 2015; Houlden et al., 1978; Lind et al., 1983; Thibaut & Walker, 1975; van den Bos et al., 1998). First, assuming that explanations are not entirely lacking in content, we conjecture, similar to A. Brown et al. (2019) and Uhde et al. (2020), that more provided information leads to higher informational fairness perceptions. Regarding effects on trustworthiness perceptions, we note that several factors contribute to a system's fairness (Colquitt & Rodell, 2015; M. K. Lee et al., 2019). Among these are *consistency* (of decision-making procedures) as well as *process and outcome control* on behalf of decision subjects (Dietvorst et al., 2018; M. K. Lee et al., 2019). *Process control* means that decision subjects have the "ability to influence what [...] data is considered by the decision maker" (M. K. Lee et al., 2019, p. 6), and *outcome control*, borrowing from Houlden et al. (1978), refers "to the ability to appeal or modify the outcome [...] once it has been made" (M. K. Lee et al., 2019, p. 6). While we do not anticipate our employed explanations to readily increase perceptions of outcome control, we conjecture that certain information may enhance assumed process control, which, in turn, affects procedural fairness perceptions (Colquitt & Rodell, 2015; M. K. Lee et al., 2019) and, ultimately, trust (van den Bos et al., 1998).

H1 As the amount of information provided increases, perceptions of informational fairness towards the ADS increase.

H2 As the amount of information provided increases, perceptions of trustworthiness towards the ADS increase.

While investigating these relationships, we are not only interested in the effects of our conditions on informational fairness and trustworthiness but also in the relationship between the latter two. Some prior work has examined the relationship

between informational fairness/justice and trust/trustworthiness in other contexts, such as Colquitt and Rodell (2011), Lance Frazier et al. (2010), and Zhu and Chen (2012). Lance Frazier et al. (2010) identified a significant positive effect of informational justice on different facets of trustworthiness perceptions in one of their two examined settings in the realm of organizational justice. Similarly, Zhu and Chen (2012), in the context of customer satisfaction in internet banking, found that informational fairness (as a component of overall systemic fairness) has a positive effect on trust. Finally, Colquitt and Rodell (2011, p. 1184) affirm that "conventional wisdom on the justice-trust connection" implies a causal path from (informational) justice to trust, and not the other way round. While these works address different use cases, we conjecture a positive relationship between informational fairness and trustworthiness perceptions for our ADS setting as well:

H3 Perceptions of informational fairness relate positively to perceptions of trustworthiness.

Experts may have a different attitude towards procedures or phenomena that touch on their area of expertise than non-experts. Slovic (1987) and Slovic et al. (1981), for instance, found differences in risk perceptions between experts and lay people. Regarding innovative (food) technologies, Siegrist (2008) notes that lay people may neither be able to assess risks nor benefits appropriately. For the specific case of ADSs, R. Wang et al. (2020) found a significant effect of computer literacy on a mix of procedural and distributive fairness perceptions; specifically, their findings suggest that fairness perceptions are lower for people with lower computer literacy. Pierson (2017), along the same lines, found that students' views on algorithmic fairness changed by increasing algorithmic literacy through lecture and discussion: students "became more likely to emphasize transparency, [and] more open to using algorithms rather than using judges" (Pierson, 2017, p. 8). Finally, intuition tells us that AI-literate people may extract more information and understanding out of ADS explanations (e.g., because they know how supervised ML in general works).

H4 People with higher AI literacy perceive an automated decision system to be more informationally fair than people with little or no knowledge in the field.

H5 People with higher AI literacy perceive an automated decision system to be more trustworthy than people with little or no knowledge in the field.

5.4 Methodology

We examine our hypotheses in the context of algorithmic lending. We argue that this is a common context that affects many people at some point in life. It is, furthermore, an area where ADSs are typically already utilized within productive settings (ACTICO, 2021; Infosys, 2019). Specifically, we confront study participants with situations where a person was denied a loan. Similar to Binns et al. (2018), we argue that, in practice, explanations are much more likely to be requested by decision subjects in response to negative outcomes; or, in other words: if someone gets the loan, interest in how and why exactly the decision was arrived at will likely drop. However, we do by no means imply that reactions to positive outcomes are unworthy of being examined—given budget constraints, we defer them to future work.

5.4.1 Study Design

We choose a between-subjects design with the following conditions: first, we reveal to study participants some basic information about the lending company. We then explain that a given individual's loan application was rejected by the company, as well as that this decision was communicated to the applying individual electronically and in a timely fashion. See Figure 5.1 on page 115 for the exact wording in our questionnaires. Afterwards, we provide one of four explanations (i.e., conditions) to each study participant. Eventually, we measure the effects of assigning different conditions—and by design of the conditions, different amounts of information—on two dependent variables: perceived informational fairness (INFF) and perceived trustworthiness (TRST) regarding the ADS. Recall that informational fairness perceptions do *not* involve an actual assessment of the system's fairness with respect to its processes or outcomes. Additionally, we measure the (self-assessed) AI literacy (AILIT) of study participants. We analyze whether differences in study participants' AI literacy affect their perceptions. All measurement items are summarized in Table 5.1 on page 116. Note that for each construct, we measure multiple items; mostly drawn (and partially adapted) from prior work.

ADS Setup The ADS for our study consists of a random forest classifier which predicts loan approval on unseen data and is able to output different explanations. For training our model, we utilize a publicly available dataset on home loan application decisions (Chatterjee, 2019), which has been used in multiple data science

Fig. 5.1.: Introduction of use case in questionnaires.

> A finance company offers loans on real estate in urban, semi-urban, and rural areas. A potential customer first applies online for a specific loan, and afterwards, the company assesses the customer's eligibility for that loan.
>
> An individual applied online for a loan at this company. The company denied the loan application. The decision to deny the loan was communicated to the applying individual electronically and in a timely fashion.

competitions on Kaggle. Note that comparable data—reflecting a given finance company's individual circumstances and approval criteria—might in practice be used to train ADSs (Infosys, 2019). The dataset at hand consists of 614 labeled (loan Y/N) observations and includes the following features: *applicant income, co-applicant income, credit history, dependents, education, gender, loan amount, loan amount term, marital status, property area, self-employment*. After removing data points with missing values, 480 observations remain, 332 of which (69.2%) involve the positive label (Y) and 148 (30.8%) the negative label (N). We used 70% of the dataset to train our ADS and use the remaining 30% as a holdout set for the experiment. After encoding and scaling the features, we trained a random forest classifier with bootstrapping (Breiman, 2001), which achieves an out-of-bag accuracy estimate of 80.1% on the held-out data. We use this classifier's predictions on the holdout set as a basis for the upcoming conditions/explanations that the study participants are confronted with. Since we are *not* asking to assess the actual (procedural or distributive) fairness of the ADS, it is not critical to quantify how fair the system really is—any such effort would be highly contestable anyhow, for reasons of incompatible fairness notions (Chouldechova, 2017; Kleinberg et al., 2017; Mulligan et al., 2019). The authors still (informally but independently) checked training data as well as output quality for any salient problems that may bias study participants' responses with respect to the dependent variables.

Explanations We impose several requirements on the explanations that we provide to study participants: overall, we employ only model-agnostic explanations (Adadi & Berrada, 2018) in a way that they could plausibly be provided to loan applicants (i.e., lay people) in real-world scenarios. While explanations can be communicated in a wide variety of ways (Adadi & Berrada, 2018; Arrieta et al., 2020), we confine ourselves to textual explanations in order to control for differences in conveyance.

Tab. 5.1.: Summary of constructs and measurement items.

Construct	Measurement items
INFF	• *The automated decision system explains decision-making procedures thoroughly.* (Colquitt & Rodell, 2015) • *The automated decision system's explanations regarding procedures are reasonable.* (Colquitt & Rodell, 2015) • *The automated decision system tailors communications to meet the applying individual's needs.* (Colquitt & Rodell, 2015) • *I understand the process by which the decision was made.* (Binns et al., 2018) • *I received sufficient information to judge whether the decision-making procedures are fair or unfair.*
TRST	• *Given the provided explanations, I trust that the automated decision system makes good-quality decisions.* (M. K. Lee, 2018) • *Based on my understanding of the decision-making procedures, I know the automated decision system is not opportunistic.* (Chiu et al., 2009) • *Based on my understanding of the decision-making procedures, I know the automated decision system is trustworthy.* (Chiu et al., 2009) • *I think I can trust the automated decision system.* (Carter & Bélanger, 2005) • *The automated decision system can be trusted to carry out the loan application decision faithfully.* (Carter & Bélanger, 2005) • *In my opinion, the automated decision system is trustworthy.* (Carter & Bélanger, 2005)
AILIT	• *How would you describe your knowledge in the field of artificial intelligence?* • *Does your current employment include working with artificial intelligence?* • *I am confident interacting with artificial intelligence.* (Wilkinson et al., 2010) • *I understand what the term* artificial intelligence *means.*

Note: All items are measured on a 5-point Likert scale and mostly drawn (and adapted) from previous studies. Recall that INFF = Informational fairness; TRST = Trustworthiness; AILIT = AI literacy.

We also pick explanations that are immediately understandable semantically—this is important so as to collect meaningful responses. On a related note, we ensure that explanations are not too long, in order to account for known issues around information overload (Bawden & Robinson, 2009). Finally, and similar to Binns et al. (2018), we pick explanations that can plausibly provide insights about a system's logic, as required, for instance, by the GDPR. Based on these preliminaries, we assign study participants to one of four conditions that involve combinations of explanations with respect to (i) factors considered by the ADS, (ii) relative importance of these factors, and (iii) counterfactual scenarios where a rejected applicant would have been granted the loan. We acknowledge that additional explanation styles would be equally interesting to consider; however, in order to keep the experiment size manageable, we must defer them to future work.

Our first condition, *(Base)*, only reveals to the study participants that the loan decision was communicated to the applying individual electronically and in a timely fashion (as in Figure 5.1 on page 115). Apart from the *(Base)* condition—which might be regarded as a black-box system—all other conditions include the additional information that the loan decision was made by an ADS (i.e., automated). The second condition, *(F)*, consists of disclosing the factors, including corresponding values for an observation (i.e., an applicant) from the holdout set whom our model denied the loan. We refer to such an observation as a *setting*. In our study, we employ two different settings in each questionnaire, where settings are chosen at random from the pool of rejected applicants. The authors, again, checked informally that no highly unusual (e.g., extreme outliers) settings were displayed that might distract study participants' perceptions and bias recorded responses.

Next, we computed permutation feature importance (Breiman, 2001) from our model and obtained the following hierarchy, using "$\succ$" as a shorthand for "is more important than": *credit history* $\succ$ *loan amount* $\succ$ *applicant income* $\succ$ *co-applicant income* $\succ$ *property area* $\succ$ *marital status* $\succ$ *dependents* $\succ$ *education* $\succ$ *loan amount term* $\succ$ *self-employment* $\succ$ *gender*. Revealing this ordered list in conjunction with *(F)* makes up our third condition, *(FFI)*. To construct our fourth condition, we conducted an online survey with 20 quantitative and qualitative researchers to ascertain which of the aforementioned factors are actionable—in a sense that people can (hypothetically) act on them in order to increase their chances of being granted a loan. According to this survey, the top-5 actionable factors are *loan amount, loan amount term, property area, applicant income, co-applicant income*. Our fourth condition *(FFICF)* is then—in conjunction with *(F)* and *(FFI)*—the provision of three counterfactual scenarios where one actionable factor each is (minimally) altered such that our model predicts a loan approval instead of a rejection. Figure 5.2

Fig. 5.2.: An exemplary setting in the *(FFICF)* condition.

Condition *(FFICF)*

A finance company offers loans on real estate in urban, semi-urban and rural areas. A potential customer first applies online for a specific loan, and afterwards the company assesses the customer's eligibility for that loan.

An individual applied online for a loan at this company. The company denied the loan application. The decision to deny the loan was made by an automated decision system and communicated to the applying individual electronically and in a timely fashion.

The automated decision system explains …

… that the following factors (in alphabetical order) on the individual were taken into account when making the loan application decision:

- Applicant Income: $3069 per month
- Co-Applicant Income: $0 per month
- Credit History: Good
- Dependents: 0
- Education: Graduate
- Gender: Male
- Loan Amount: $71,000
- Loan Amount Term: 480 months
- Married: No
- Property Area: Urban
- Self-Employed: No

… that different factors are of different importance in the decision. The following list shows the order of factor importance, from most important to least important:

Credit History ≻ Loan Amount ≻ Applicant Income ≻ Co-Applicant Income ≻ Property Area ≻ Married ≻ Dependents ≻ Education ≻ Loan Amount Term ≻ Self-Employed ≻ Gender

… that the individual would have been granted the loan if—everything else unchanged—one of the following hypothetical scenarios had been true:

- The co-applicant income had been at least $800 per month
- The loan amount term had been 408 months or less
- The property area had been rural

on page 118 shows one exemplary setting, including introduction of use case and explanations in the *(FFICF)* condition. Our four conditions are summarized as follows:

(Base)	Baseline without further explanations
(F)	Disclosure of factors
(FFI)	Disclosure of factors and factor importance
(FFICF)	Disclosure of factors, factor importance, and counterfactuals

Note that the order of provided explanations *(Base)*→*(F)*→*(FFI)*→*(FFICF)* is not arbitrary: each subsequent condition provides the exact same information as the previous one *and more*. Since, for instance, factor importances implicitly reveal which factors the ADS considers, this would not necessarily hold true for *(FI)*→*(FIF)*.

5.4.2 Data Collection

Study participants for our online study were voluntarily recruited via Prolific (Palan & Schitter, 2018) and asked to rate their agreement with multiple statements regarding our dependent variables as well as their AI literacy on 5-point Likert scales, where 1 corresponds to "strongly disagree" and 5 denotes "strongly agree." Additionally, we included multiple open-ended questions in the questionnaires to be able to better understand the reasoning behind study participants' quantitative responses. The study participants were randomly and in equal proportions assigned to one of the four conditions, and each participant was provided with two consecutive questionnaires associated with two different settings. We collected 401 responses, of which 4 had to be eliminated due to failure to pass one or more attention checks. Thus, we obtained 397 analyzable responses. Among the study participants, 60% indicated to be male, 39% female, and the remaining participants either responded with "non-binary" or chose not to disclose their gender; 46% were students, 27% employed full-time, 8% employed part-time, 7% self-employed, 11% unemployed, less than 1% retired, and 1% chose not to disclose their profession. The reported average age was 25.7. Study participants were monetarily compensated above the recommended minimum pay of $6.50 per hour.

5.5 Quantitative Analyses and Results

We now examine the effects of our conditions and people's (self-assessed) AI literacy on perceived informational fairness and trustworthiness of our ADS. First, we establish our measurement model, describing a confirmatory factor analysis and reporting correlations and factor loadings. After that, we present the results of group difference analyses for our conditions with tests for pairwise comparison. Finally, we report our findings on the validation of our hypotheses **H1** to **H5** with a full structural equation model.

5.5.1 Measurement Model

In order to assess the validity and the reliability of our constructs, we conduct a confirmatory factor analysis and assess the results with respect to multiple measures. As measures for convergent reliability, we examine the average variance extracted (AVE) and composite reliability (CR). For the constructs of informational fairness and trustworthiness, AVE is above the recommended threshold of 0.5, whereas the AVE of AI literacy is 0.41. According to Fornell and Larcker (1981), if AVE is low, convergent validity of a construct can still be sufficient if CR is above 0.6, which is the case for all three constructs, including AI literacy (see Table 5.2 on page 121). In fact, the CR of our three main constructs, informational fairness (0.88), trustworthiness (0.94), and AI literacy (0.72) is above the recommended threshold of 0.7 (Barclay et al., 1995), indicating that our convergent validity is adequate for AI literacy as well, despite the lower AVE measure.

Cronbach's alpha (CA) values for our constructs are larger than the recommended threshold of 0.7, thus showing good reliability for all constructs (Cortina, 1993). Validity and reliability measures are summarized in Table 5.2 on page 121. Our matrix of factor loadings, demonstrated in Table 5.3 on page 121, shows that all items load highly (> 0.5) on one factor each with low cross-loadings, and the correlations between factors are all below 0.7 (see Table 5.2 on page 121). Furthermore, the AVE value of each of our constructs is larger than the squared correlation of that construct with every other construct, which is a discriminant validity measure suggested by Chin (1998) and Fornell and Larcker (1981). Therefore, convergent validity and discriminant validity are sufficiently satisfied. We test for multicollinearity by determining the variance inflation factors (VIF). According to a rule of thumb, the VIF has to be lower than 10, otherwise, multicollinearity might be a serious

problem (Vittinghoff et al., 2011). All VIFs in our model are less than 2, which indicates that there are no issues of multicollinearity.

Tab. 5.2.: Correlations and measurement information for latent constructs.

Construct	M	SD	CA	CR	AVE	INFF	TRST	AILIT
INFF	3.15	0.87	0.87	0.88	0.60	1.00		
TRST	3.26	0.84	0.94	0.94	0.73	0.67	1.00	
AILIT	2.87	0.61	0.71	0.72	0.41	0.25	0.18	1.00

Note: M = Mean; SD = Standard deviation; CA = Cronbach's alpha; CR = Composite reliability; AVE = Average variance extracted.

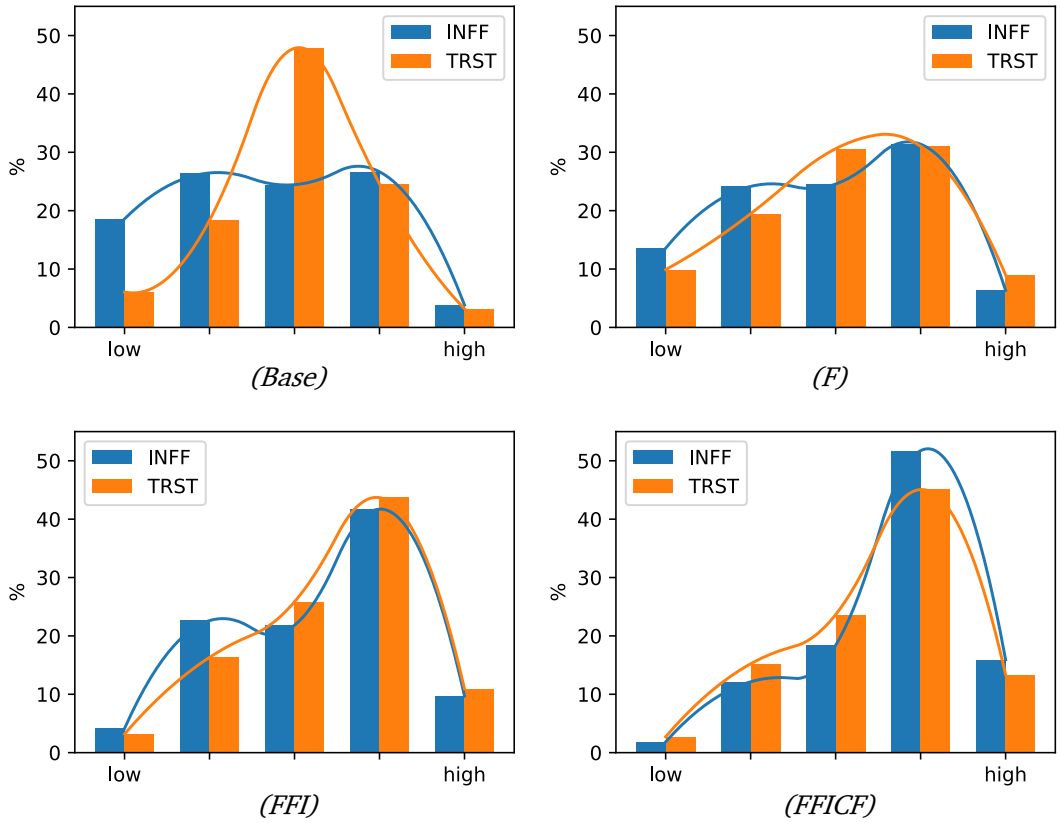
Tab. 5.3.: Standardized loadings of measurement items on constructs.

Measurement item	INFF	TRST	AILIT
INFF1	**0.95**	−0.11	−0.03
INFF2	**0.65**	0.21	0.01
INFF3	**0.52**	0.10	0.05
INFF4	**0.79**	0.01	0.03
INFF5	**0.76**	0.01	0.00
TRST1	0.24	**0.66**	−0.05
TRST2	0.20	**0.51**	−0.08
TRST3	0.01	**0.90**	−0.01
TRST4	−0.08	**0.97**	0.06
TRST5	0.02	**0.90**	0.05
TRST6	−0.09	**1.01**	0.00
AILIT1	0.08	−0.11	**0.73**
AILIT2	0.06	−0.03	**0.53**
AILIT3	−0.12	0.17	**0.67**
AILIT4	0.00	−0.02	**0.58**

5.5.2 Analysis of Group Differences

Since we cannot confirm the assumption of normality for all variables, we conduct multiple nonparametric Kruskal-Wallis H tests for multiple group comparisons (Kruskal & Wallis, 1952). Afterwards, we carry out pairwise comparisons using Bonferroni-corrected Mann-Whitney U tests (Mann & Whitney, 1947). With these tests, we initially assess the effects of our four conditions revealing different amounts of information (AMTIN) on the constructs of informational fairness (INFF) and trustworthiness (TRST). Overall, we find a significant effect between different conditions

Fig. 5.3.: Distributions of responses for informational fairness (INFF) and trustworthiness (TRST) per condition.



on perceptions of informational fairness ($p < 0.001$) as well as on perceptions of trustworthiness ($p < 0.001$). A Mann-Whitney U test for pairwise comparisons shows that the effect for informational fairness is significant ($p < 0.05$) between all conditions except *(Base)* and *(F)*. The effect for trustworthiness is significant between *(Base)* and *(FFI)*, *(Base)* and *(FFICF)*, as well as *(F)* and *(FFICF)*, and marginally significant between *(F)* and *(FFI)* ($p = 0.052$). Looking at the mean response values for INFF and TRST by condition (see Table 5.4 on page 123), we note that they are increasing as more information is shown to study participants. Please refer to Figure 5.3 for the distribution of responses by condition, and to Table 5.5 on page 123 for a detailed summary of the results of the Mann-Whitney U tests.

5.5.3 Hypotheses Testing

We estimate a full structural equation model (SEM), the results of which are depicted in Figure 5.4 on page 124. We also report more exhaustive information, including

Tab. 5.4.: Means and standard deviations of response values for informational fairness (INFF) and trustworthiness (TRST) by condition.

Condition	M(INFF)	SD(INFF)	M(TRST)	SD(TRST)
(Base)	2.71	1.16	3.01	0.89
(F)	2.93	1.16	3.10	1.12
(FFI)	3.30	1.05	3.43	0.99
(FFICF)	3.68	0.94	3.51	0.99

Note: All items were measured on 5-point Likert scales.

Tab. 5.5.: Pairwise differences in perceptions of informational fairness (INFF) and trustworthiness (TRST) between conditions.

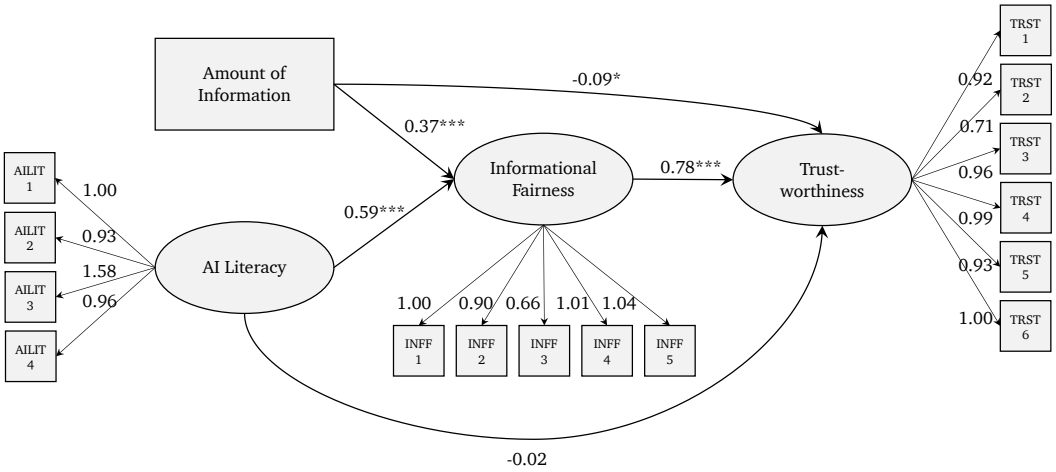
INFF			TRST		
Condition 1	**Condition 2**	**Difference**	**Condition 1**	**Condition 2**	**Difference**
(Base)	*(F)*	n/s	*(Base)*	*(F)*	n/s
(Base)	*(FFI)*	***	*(Base)*	*(FFI)*	***
(Base)	*(FFICF)*	***	*(Base)*	*(FFICF)*	***
(F)	*(FFI)*	*	*(F)*	*(FFI)*	n/s
(F)	*(FFICF)*	***	*(F)*	*(FFICF)*	**
(FFI)	*(FFICF)*	**	*(FFI)*	*(FFICF)*	n/s

*Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; n/s: not significant.*

standard errors, z-values, p-values, and standardized path estimates in Table 5.6 on page 125. Consistent with using Kruskal-Wallis H tests for group comparisons, we estimate our SEM using unweighted least squares (ULS) because this estimator makes no distributional assumptions. We assess the fit of our model with multiple common measures: the comparative fit index (CFI) as well as Tucker-Lewis index (TLI) should be above 0.9 (Kline, 2015), root mean square error of approximation (RMSEA) below 0.05 (Browne & Cudeck, 1992), and standardized root mean square residual (SRMR) below 0.08 (Hair Jr. et al., 2016) to indicate good model fit. Our model's values are: 0.997 (CFI); 0.997 (TLI); 0.024 (RMSEA); 0.051 (SRMR). Hence, all considered fit measures meet the required thresholds. Note that the chi-square test is not a meaningful measure of model fit in our case because variables are not normally distributed, and because we apply the ULS method to estimate our model (Kenny, 2015).

In the following, we use a shorthand for our variables: AMTIN, AILIT, INFF, TRST (as introduced in Section 5.4.1). To investigate our hypotheses, we first examine the effect of AMTIN on INFF. As expected, and previously supported by the Kruskal-

Fig. 5.4.: Full structural equation model (SEM) including measurement model.



Note: $*p < 0.05$*; ***p < 0.01*; ****p < 0.001$.

Wallis H test as well as the comparison of means between different conditions, increasing AMTIN has a significant positive effect on INFF ($0.37***$). Hence, **H1** is supported.

Next, we examine the influence of AMTIN on TRST. The results of the Kruskal-Wallis H test from Section 5.5.2 indicate that there is a significant positive relationship between AMTIN and TRST. However, a mediation analysis within the SEM reveals that this effect is mediated by INFF. When assessing this mediating effect more closely in the context of our SEM, a small direct effect of AMTIN on TRST persists. Interestingly, in the context of the model, the stronger effect of AMTIN on TRST through INFF is positive, while the smaller but significant remaining direct effect is negative ($-0.09*$). We discuss this in more detail in Section 5.7. Overall, **H2**, which conjectures a positive *total* (i.e., direct plus indirect) effect of AMTIN on TRST, is supported in our study.

The SEM's path coefficient concerning **H3** ($0.78***$) confirms that there is a statistically significant positive relationship between INFF and TRST, which confirms **H3**. This result provides a crucial individual piece of information in the context of the analysis of INFF as a mediator between AMTIN and TRST. As presumed in **H4**, the path coefficient between AILIT and INFF ($0.59***$) confirms the conjecture of a significant positive relationship between these two variables. Therefore, **H4** is supported by our results. Similar to our findings with respect to the effect of AMTIN on TRST, the relationship between AILIT and TRST is also mediated by INFF. The analysis of effects within the full SEM confirms a strong indirect effect of AILIT on TRST through INFF, but the remaining direct effect of AILIT on TRST is not

Tab. 5.6.: Results of model estimation.

Path	Estimate	SE	z-value	p-value	Std.lv
AILIT → INFF	0.59^{***}	0.08	7.01	< 0.001	0.31
AMTIN → INFF	0.37^{***}	0.03	14.25	< 0.001	0.47
INFF → TRST	0.78^{***}	0.05	15.30	< 0.001	0.78
AILIT → TRST	-0.02	0.07	-0.24	0.81	-0.01
AMTIN → TRST	-0.09^{*}	0.04	-2.55	0.01	-0.11

Note: $^{*}p < 0.05$*;* $^{**}p < 0.01$*;* $^{***}p < 0.001$.

significant. Hence, the effect of AILIT on TRST is completely mediated by INFF. In conclusion, **H5**, which assumes a positive relationship between AILIT and TRST, is supported.

5.6 Qualitative Analysis

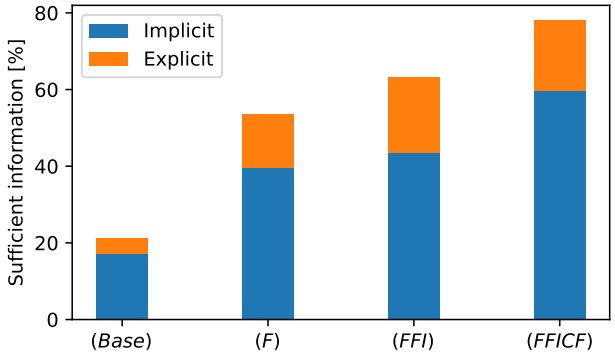
In this section, we aim to understand people's perceptions in more detail. To that end, we collected responses to open-ended questions regarding (i) what information study participants think they are missing (if any) to be able to judge whether the system behaves fairly, and (ii) study participants' perceptions of (in)appropriateness of the given explanations. These questions were part of each condition. The authors jointly coded the qualitative data inspired by grounded theory (Charmaz, 2003), that is, codes evolved as we analyzed the data. In total, 982 text passages were coded over five coding sessions with `MAXQDA` (Kuckartz & Rädiker, 2019). The emerging themes from the collected responses are summarized in the following subsections. Every direct quote is provided with a unique identifier, introduced with the "#" symbol. Some responses contain statements with respect to multiple themes; hence, percentages do not always add up to 100%.

5.6.1 What Information Is Missing?

For this question, we coded 421 text passages from study participants' responses to the open-ended question: *If you don't feel you received sufficient information to judge whether the decision-making procedures are fair or unfair, what information is missing?* We distinguish responses by condition and examine how many study participants felt that they received sufficient information (either by saying so explicitly

or by not answering this question altogether). The latter is visually summarized in Figure 5.5.

Fig. 5.5.: Percentage of responses indicating that study participants received sufficient information to judge whether the system's procedures are fair or unfair.



Note: We capture both cases where study participants indicated explicitly *in their responses that they received sufficient information, or* implicitly *by not answering the respective question.*

***(Base)* Condition** Most study participants (79%) assigned to this condition felt that they did not receive sufficient information; 17% did not answer the question, and 4% explicitly stated that they are not missing any information. Little surprisingly, when asked which information they are missing, study participants were interested in knowing why the system made particular decisions; 37% of all responses contained statements substantially similar to this: *"All I know is that the loan was denied and not the reason why"* (#1315). Similarly, 30% of responses inquired about decision criteria that underlie the rejected loans: *"I have no way to know what references the company may or may not use to consolidate a decision about the eligibility of an individual for a particular loan, and therefore I might or might not find the procedures to be truly fair"* (#1260). 16% of responses also thought that decision-making procedures in general must be explained more thoroughly, arguing that *"everything to do with how they made their decision of whether to accept the loan or not [is missing]"* (#1234). Some study participants were more specific as to what explanations they need: 18% indicated that relevant factors of applicants would be helpful to know (#1259: *"To decide whether the decision-making procedures are fair or unfair, I probably would need to know how the client was economically and other factors such as criminal records"*); and 6% of responses requested counterfactual-type insights related to recourse, e.g., *"what he can do to try again"* (#1265).

***(F)* Condition** In the condition where decision-making factors were shown, already 54% of study participants indicated that they received sufficient information. Of those who indicated that more information is needed, 15% are still interested in the *why* behind the rejections (#587: *"I think clearly spelled reason is missing instead of numbers"*). 15% still thought that more information with respect to decision criteria is needed. Interestingly, knowing what factors are used by the ADS raises further, more specific, questions as to why (i) these given factors are considered (#731: *"There needs to be more in depth explanations given as to why these factors are taken into consideration"*), and (ii) not others, for instance, *"how many loans have they taken out in the past, what is the money going to be spent on etc"* (#663). Overall, 23% of study participants requested these justifications. Another 10% of responses indicated that it would be necessary to know how each factor impacts the final decision—both in terms of weighting (#474: *"What kind of value does each factor hold?"*) and monotonic relationships with the outcome (#602: *"The factors are told, but not which ones influenced the response positively of negatively."*) Finally, 3% are interested in counterfactual explanations, for instance, *"how the factors should differ for the application to be approved"* (#474).

***(FFI)* Condition** In this condition, only 37% of study participants requested further information. Among these, 15% still requested more information with respect to reasons why the ADS rejected the applications; and 17% felt that they still had not received sufficient information regarding decision-criteria (#677: *"There is not enough information about what thresholds have to be met to qualify for a loan."*) On a related note, 6% of study participants wanted to see more explanation as to why *"the [factor importance] ranking is the way it is"* (#764). Similar to the *(F)* condition, some study participants (10%) wanted to know why certain factors of the applicants are not being considered by the ADS. 3% of study participants still needed to know how exactly specific factors impact the final decision (#684: *"I don't know the significance level/weight assigned to [the factors]"*); and another 3% specifically requested counterfactual-type explanations. A newly occurring theme is with respect to communication of the explanations, as 3% requested *"less formal descriptions"* (#714).

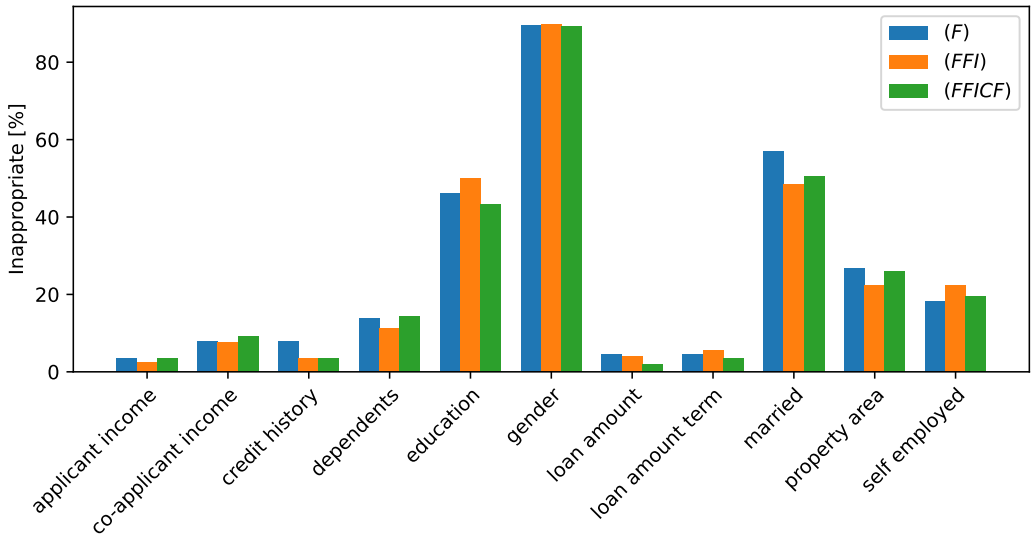
***(FFICF)* Condition** In our condition with the highest amount of provided information, only 22% requested additional information. Generally speaking, responses are more dispersed compared to other conditions. Some study participants still alluded to missing justification with respect to the given selection and importance of relevant factors (overall 14%), and others (7%) still asked for more information

on the relationship between certain input factors and the outcome (#796: *"Since I think gender being a factor is unfair, not knowing the degree to which it affects the outcome seems to be a deficiency."*) 6% of study participants were interested in the rationale behind providing given counterfactuals: *"The factors that could have changed the outcome [are revealed], but not the reason why those [...] factors would be needed. Ex.: Why would a rural area be more easily accepted?"* (#856). Interestingly, no study participant requested additional information as to why the ADS rejected the applicants—as opposed to the other conditions. Yet, 11% still requested more information with respect to decision criteria, for instance, *"the thresholds that are required for a loan to be accepted"* (#800). 6% stated that processes were generally still not fully clear; however, some acknowledged that this might not necessarily be expedient in the first place (#863: *"It's not clear how practically the priority system works, but I can understand it would be too hard to explain, and probably most of the people wouldn't understand it anyway."*)

5.6.2 Appropriateness of Individual Explanations

We also asked study participants about their feelings of (in)appropriateness of isolated explanations, specific to the condition they were assigned to: *Why do you think {some factors, the order of factor importance, some counterfactual scenarios} are appropriate or inappropriate?* For that, we coded 561 text passages and summarized the main themes for each type of explanation.

Fig. 5.6.: Inappropriate factors according to responses from study participants, broken down by condition.



Factors Only 14% of responses explicitly stated that (at least a subset of) the factors considered by the ADS were appropriate—mostly those related to an applicant's financial situation (#602: *"Economic factors seem apropriate* [sic] *to me. Self employment sometimes involves risks and it is a relevant factor also."*) We also asked study participants to check specific factors they deem inappropriate—this is visualized (by condition) in Figure 5.6 on page 128. Among responses with respect to inappropriate factors, two general themes emerged: 72% indicated that some factors are (causally) irrelevant for deciding on creditworthiness (#632: *"Some of the more social-oriented factors (ie education, gender, dependents) aren't necessarily indicative of someone's ability to pay back a loan"*), and 28% found the usage of certain factors (primarily *gender, education*, and *married*) morally wrong (#561: *"In the world we live, i dont* [sic] *think gender is something to even be at question, neither marriage."*) Interestingly, study participants often assumed that the sheer presence of a factor like *gender* means that it is being used with malicious intent: *"Gender can be somewhat problematic because all people deserve to have the right to the loan and not only men"* (#637), or, *"some factors like gender are plain racist to make a financial decision"* (#647).

Factor Importance Generally speaking, most study participants found the order of factor importance reasonably appropriate. Many responses resembled this: *"I may not agree with the placement of every single factor, but overall i think they are ranked appropriately"* (#695). Yet, 35% still suggested concrete changes with respect to the order of importance; particularly around assigning less weight to *education* and *marital status*. 14% were still entirely put off by the fact that *gender* or *marital status* were used in the decision-making process. However, learning that *gender* is the least important factor made many study participants feel better with respect to appropriateness of procedures (#510: *"It is appropriate. Gender should be considered the least and credit history is most important."*) One study participant even suggested that *"gender could play a part in the decision making, but not a big one so it's good as it is"* (#751). (Recall that *gender* was ranked last in our explanation, see Section 5.4.1.)

Counterfactual Scenarios 47% of coded responses indicated that the provided counterfactual scenarios are appropriate, for instance, endorsing that they *"are all financial and based on the ability of the loan to be paid back"* (#448). However, 20% questioned the effectiveness of adhering to some of the counterfactual recommendations; especially regarding suggested changes to *co-applicant income* or *property area*: *"These factors do not change the fact that an applicant can or can not pay his/her debt"*

(#454). *Actionability* of counterfactual scenarios was another important theme: 9% overall addressed this, being appreciative that some counterfactual scenarios are explicitly actionable (#836: *"Changing the loan term is possible immediately"*) and disenchanted when not (#462: *"Some hardly achivable* [sic] *scenarios must be met to ensure the bank [will] be repayed* [sic]*."*) Some themes were addressed by fewer study participants but are highly interesting: one study participant was confused by the "direction" of suggested changes: *"Instead of a short loan amount term, it could be a bit longer"* (#778). Others were seemingly distracted by suggested changes that are (too) small: *"The incomes are so close to the required that it shouldn't matter"* (#447). Finally, some study participants hinted at potential inconsistencies between individual explanations: *"It seems odd that loan amount term is placed so low when it was one of the areas the individual could change to obtain the loan"* (#435).

5.7 Discussion and Implications

In this section, we link our quantitative results to qualitative insights to get a better understanding as to why certain effects were observed, and we analyze and discuss in more detail the findings from the fitted SEM. Finally, we allude to several implications of our work.

Connecting Quantitative and Qualitative Findings We have seen that both perceptions of informational fairness and trustworthiness increase as more explanations are provided to study participants—however, INFF at a much higher rate than TRST. Interestingly, many study participants in the *(Base)* condition, who do not receive any further explanations with respect to the inner workings of the ADS, do *not* find this black-box system to be overly problematic with respect to informational fairness: as can be seen in Figure 5.3 on page 122, study participants' responses for INFF are approximately equally distributed across ratings 1–4. This might be due to people's expectations; one study participant simply stated that this *"seems to be standard practice"* (#1212) in terms of explaining ADSs. From Table 5.5 on page 123, we infer that providing relevant factors *(F)* to study participants does not significantly increase INFF. A likely reason for this observation is that study participants asked for significant follow-up information with respect to *how* the factors are used for decision-making. Both the differences for *(F)*→*(FFI)* and *(FFI)*→*(FFICF)* are significant for INFF. Considering the qualitative findings from Section 5.6.1, this seems little surprising as the complementary explanations (e.g., factor importance in *(FFI)* over *(F)*) were specifically requested by study participants.

While some explanations clearly helped study participants understand the given ADS better, they also reveal certain aspects that might be detrimental to people's trust. Similar to INFF, one might have expected to see lower ratings for TRST in the *(Base)* condition. Instead, study participants' responses for TRST are symmetrically distributed around the mean of 3 (see Figure 5.3 on page 122). Regarding marginal effects of explanations on TRST, we note that none of *(Base)*→*(F)*, *(F)*→*(FFI)*, or *(FFI)*→*(FFICF)* lead to statistically significant changes in study participants' perceptions. As for *(Base)*→*(F)*, study participants' trust appears to be hampered by the experience that certain (presumably) inappropriate factors (e.g., *gender*) are being considered by the ADS. While the change *(F)*→*(FFI)* is marginally significant ($p = 0.052$) for TRST, we still suspect a certain attenuation due to study participants' disagreement with the relative importance ranking of certain factors like *education* and *married*. On the other hand, from analyzing the qualitative statements, we might assume *gender* playing the least important role in the decision-making process had a positive effect on study participants' trust. As for *(FFI)*→*(FFICF)*, we suspect that a potential positive effect of counterfactual explanations on perceived outcome control (Houlden et al., 1978) might have been overshadowed by the fact that several study participants found some of the provided scenarios incomprehensible, ineffective, or unactionable.

Interpreting SEM Results In addition to confirming significant *total* effects (see Figure 5.4 on page 124) of the amount of information (AMTIN) on INFF (0.37***) and TRST ($0.37 \cdot 0.78 - 0.09 = 0.20$***), we also learn that study participants' (self-assessed) AI literacy (AILIT) is strongly related to INFF (0.59***) and TRST (0.44***), implying that we observe higher INFF and TRST ratings for higher AI-literacy people—given our study setup. Additionally, we see a strong positive relationship between INFF and TRST (0.78***). The SEM also lets us decompose total effects of AMTIN and AILIT on TRST into direct and indirect (through the mediator INFF) effects. This is shown in Table 5.7 on page 132. We see, for instance, that the direct effect of AILIT on TRST (-0.02) is not significantly different from zero when INFF is acting as a mediator. Since the indirect effect AILIT→INFF→TRST is significantly positive (0.46***), we observe a complete mediation of the effect of AILIT on TRST through INFF. A similar observation can be made for the effect of AMTIN on TRST: the total effect consists of a significantly *positive* indirect effect through INFF (0.29***) as well as a small *negative* direct effect (-0.09*). Hence, we conclude that increasing AMTIN does *not directly* increase TRST, but that the positive total effect stems from the strong indirect effect through INFF. This phenomenon is sometimes also referred to as *inconsistent mediation* (Kenny, 2015; MacKinnon et al.,

2007). Future work should further investigate the link between INFF and TRST for other scenarios.

Tab. 5.7.: Decomposition of effects on perceived trustworthiness.

	Direct effect	Indirect effect	Total effect
AMTIN on TRST	-0.09*	$0.37 \cdot 0.78 = 0.29$***	0.20***
AILIT on TRST	-0.02	$0.59 \cdot 0.78 = 0.46$***	0.44***

Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Implications Our work has several implications for the design of automated decision systems and explanations thereof. Revealing to (potential) decision subjects *what* information about them is used and *how* exactly individual factors affect the outcome is something that appears to go a long way towards facilitating informational fairness. We have also seen that many people require an understanding of (assumed) monotonic relationships between individual features and outcome (#856: *"We don't know if being married is a good or bad thing in this case."*) However, these types of global monotonic relationships cannot generally be derived from nonlinear ML models—something that has been discussed, among others, by Rudin (2019) and S. Wang and Gupta (2020). Employing inherently interpretable (e.g., linear) models might be a potential remedy.

We made a similar observation with respect to monotonicity for counterfactual explanations: people are put off when the "direction" of suggested change(s) contradicts commonly-held assumptions (e.g., if a *decrease* in income were suggested in order to get the loan). System designers must therefore pay close attention that counterfactual scenarios or general recommendations on recourse are intuitive, meaningful, and actionable. Regarding the latter, we have observed that certain factors are deemed actionable by some study participants and immutable by others. This poses further challenges with respect to individualizing explanations (Kühl et al., 2020); this is also relevant for people with different AI backgrounds as their perceptions differ. In general, however, counterfactual explanations appear to be effective in a way that they help people understand *"where [an] applicant fell short"* (#731).

From the analysis of qualitative data (also confirmed quantitatively), we learned that study participants in the *(Base)* condition specifically requested explanations related to both factor importance and recourse, including why the ADS decided negatively. This suggests the employment of both explanation types in a complementary fashion. Designers will have to ensure, however, that they are *consistent* with one another.

For instance, people seem to expect that recommendations for recourse (e.g., that income should be increased) apply to the factors that are most important in the decision-making process. Since individual explanations are often automatically and independently generated, this poses a significant technical challenge. Our findings also suggest that informational fairness might be further increased by providing rejected loan applicants with a crisp statement in lay people's terms as to why they were denied. Finally, regarding the usage of sensitive information like *gender*, it should be clearly justified why and how (if at all) this information is used, and that this is not automatically to the disadvantage of marginalized groups; for instance, in the case of affirmative action (Holzer & Neumark, 2000).

5.8 Limitations and Outlook

We acknowledge limitations of our work that open up avenues for future studies. Firstly, we investigated only one setting where ADSs are currently used to inform consequential decisions: lending. Our study design should be replicated and the results should be compared in different settings, for instance, hiring or university admissions, where the relevant factors will be significantly different. It would also be interesting to work with domain experts, as opposed to crowdworkers. Future work should further examine the complementarity and interplay of other explanation styles (e.g., case-based or demographic explanations (Binns et al., 2018)). Furthermore, our quantitative results (including SEM) are contingent upon the concrete instantiation of our ADS including the employed explanations, which limits our ability to generalize findings.

While we informally checked the model as well as the underlying data and all derived explanations so as to ensure behavior that might be representative of many real-world applications, it would be insightful to randomize different aspects about the model's quality and compare the results. More specifically, if we managed to construct—broadly speaking—a trustworthy ADS and an untrustworthy ADS, we would be able to contrast people's perceptions for either system. This would allow to derive insights with respect to *(un)warranted* perceptions, that is, (i) are people *actually* able to spot problematic behavior of ADSs, and (ii) do they trust the system if and only if the system is trustworthy? In fact, for an untrustworthy ADS, we would ideally expect that more explanations lead to *higher* informational fairness perceptions but to *lower* trust. If perceptions of trustworthiness increase regardless of the actual trustworthiness of the ADS, this would indicate serious issues around

over-reliance (Skitka et al., 2000) or automation bias (De-Arteaga et al., 2020; Goddard et al., 2014), and must be avoided by system designers at all costs.

We also acknowledge that our work does not explicitly take into account potential issues around information overload (Bawden & Robinson, 2009): while we specifically examine situations where selected explanations convey complementary information, unsystematic provision of more and more explanations will likely have undesirable effects. The authors suggest by no means that more information is always better. Finally, we hope that this work can serve as a stepping stone for further empirical research on the complementarity and interplay of different explanations and their effects on people's perceptions towards AI systems.

6 Comparing Fairness Perceptions Towards AI-Based Versus Human-Based Decisions

In this chapter, we conduct a second mixed-method online study to discern variations in human perceptions when the final decision is made by an artificial intelligence (AI) system versus a human. Interestingly, our findings suggest that people perceive AI systems as fairer than human decision makers. Our analyses also indicate that an individual's AI literacy influences their perceptions, with those possessing higher AI literacy showing a stronger preference for AI systems over human decision makers. Conversely, individuals with lower AI literacy do not exhibit significant differences in their perceptions. From our qualitative analyses, we infer that the preference for automation often arises from people's belief in the (assumed) objectivity and bias-free nature of AI systems. This raises questions about the calibration of human perceptions.

6.1 Introduction

Over the recent years, a considerable amount of scholarly effort has been channeled towards identifying and rectifying instances of unfairness in automated decision systems (ADSs). However, a substantial share of this research has been primarily

This chapter is based on published work. To enhance the reading experience and maintain overall consistency of the thesis, we removed the abstract and made several minor adjustments. The original paper can be accessed via:

Schöffer, J., Machowski, Y. & Kühl, N. (2022). Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making. *Proceedings of the 55th Hawaii International Conference on System Sciences 2022* (pp. 1095–1102). http://hdl.handle.net/10125/79466.

centered around defining the notion of fairness and adapting machine learning (ML) algorithms to adhere to various statistical parity constraints. This focus largely overlooks the invaluable input from individuals who are directly impacted by automated decisions. Among others, Srivastava et al. (2019) emphasize this need for better understanding people's attitudes towards fairness of ADSs. This work is vital not only from a moral perspective but also regarding the effective design and implementation of ADSs—with the end goal of creating systems that are fair, trustworthy, and, as a result, suitable for wide adoption. To that end, we conduct a mixed-method study to better understand people's perceptions of fairness and trustworthiness towards ADSs in comparison to the (hypothetical) scenario where a human instead of an ADS makes the decision. We furthermore analyze how these perceptions may change depending on people's background and experience with AI.

It is widely understood that opaque (i.e., black-box) ML models do not allow for meaningful interpretations as to how or why certain outcomes were arrived at (Peters et al., 2020; Wanner et al., 2020). Prior research has also shown that explanations can be an effective tool for more transparent decision-making (Adadi & Berrada, 2018; Meske et al., 2022). Therefore, in this work, we provide study participants with thorough explanations regarding decisions—identical for both the case of the ADS and the human decision maker. The context of our study is lending, which is a common high-stakes application of ADSs (ACTICO, 2021).

6.2 Background and Related Work

Harris and Davenport (2005) characterize *automated decision systems* as those designed to minimize human involvement in decision-making processes. Given their escalating utilization across various critical sectors (Grote & Berens, 2020; Kuncel et al., 2014; Townson, 2020), it is imperative to ensure that these ADSs deliver decisions that are both fair and transparent. However, recent history has seen numerous instances where algorithms have exhibited biased decision-making based on factors such as gender or race (Angwin et al., 2016; Buolamwini & Gebru, 2018; Heaven, 2020). Furthermore, the ML models underpinning these systems are often perceived as black boxes, thereby rendering their interpretation a significant challenge (Arrieta et al., 2020; Wanner et al., 2020).

6.2.1 Explainable AI

The importance of *explainable AI* keeps rising as increasingly sophisticated (and opaque) AI systems are used to inform increasingly high-stakes decisions. Explanations can be distinguished along different dimensions. Adadi and Berrada (2018), for instance, differentiate between model-specific and model-agnostic explanations. *Model-agnostic explanations* refer to methods that are not bound to a single type of ML model and are therefore more generalizable—which is why we employ them in this work. Examples of the model-agnostic (example-based) explanation style, which provide information people can potentially act upon, are *counterfactual explanations* (Fernandez et al., 2019). In brief, counterfactual explanations provide people with information regarding the minimum changes that would lead to an alternative (generally the desirable) decision. Meske et al. (2022) discuss different types of explanations relevant in information systems research, particularly model-agnostic explanations. They argue that transparency is essential for evaluating automated systems. People affected by an automated decision may be particularly interested in explanations to assess the fairness or trustworthiness of the associated ADS. Other popular model-agnostic explanation styles include the provision of the relevant features used by an ML model or (permutation) feature importance (Breiman, 2001), both of which we employ in this work since they could be plausibly provided by both human and automated decision makers. We refer to Adadi and Berrada (2018) and Goebel et al. (2018) for more in-depth literature on the topic of explainable AI.

6.2.2 Perceptions of Fairness and Trustworthiness

Scholars in AI and human-computer interaction (HCI) have begun to turn their attention towards the examination of perceived fairness and trustworthiness within the sphere of automated decision-making. Studies as the ones by Binns et al. (2018) and Dodge et al. (2019) have conducted comparative analyses of fairness perceptions, notably in relation to various styles of explanation. Their work highlights the differences in the effectiveness of individual explanation strategies, while also acknowledging the absence of a single, superior approach for explaining automated decisions.

In a similar context, M. K. Lee (2018) delves into the comparison of fairness and trust perceptions in scenarios where decisions are made by either humans or an algorithm, under the umbrella of algorithmic management. This includes tasks like work scheduling and evaluations. The research findings suggest that automated decisions are often perceived as less fair and trustworthy when these encompass

tasks typically undertaken by humans. Additional exploration by M. K. Lee and Baykal (2017) compares the perception of algorithmic decisions with decisions made by a group. A particularly intriguing finding from M. K. Lee et al. (2019) is the potential deterioration in the perception of fairness for individuals who understand the working of the algorithm, especially when their personal fairness concepts deviate from those embedded in the algorithm. From the trustworthiness perspective, Kizilcec (2016) underscores the importance of achieving an optimal balance of transparency to foster trust in automated decision systems.

6.2.3 Human Versus Automated Decisions

Individuals encounter algorithms and automation in a variety of contexts. Consequently, it becomes vital to understand how this form of automation influences people emotionally and to deduce the social impact of algorithms. There exists a dichotomy where engineers often display optimism about the capability of ADSs to identify and mitigate human biases and stereotypes, while laypeople typically exhibit concerns about potential AI dominance (Crawford & Calo, 2016). However, research by Castelo et al. (2019) reveals that when decisions are perceived as objective, individuals lean towards automated advice, whereas for subjective decisions, they prefer advice from humans. This insight aligns with the findings by M. K. Lee (2018). Moreover, Castelo et al. (2019) suggest that the perceived objectivity of a task can be modified, subsequently increasing the trust and reliance placed on automated decisions. Perhaps less surprisingly, Kramer et al. (2018) have found that individuals' preferences for human or AI-based decisions are also influenced by their previous experiences with ADSs.

A prominent issue with ADSs is that individuals are often oblivious of their existence. For instance, Eslami et al. (2015) shed light on people's unawareness of the algorithm steering Facebook's news feed. Over half of the participants in their study were ignorant of the algorithm's influence, with some expressing anger and dissatisfaction in response. This lack of awareness, coupled with negative experiences (Buolamwini & Gebru, 2018; Satariano, 2020), among other factors, could be contributing to the profound aversion that many individuals have towards algorithms (Dietvorst et al., 2015; Edelman, 2021). Elevating people's awareness of ADSs, for instance, by proactively disclosing the nature of the decision maker, and considering their perceptions of these systems, may help enhance acceptance in situations where ADSs can deliver better and fairer decisions than humans.

6.2.4 Our Contribution

We aim to complement prior research to better understand people's perceptions of fairness and trustworthiness towards ADSs versus human decision makers in high-stakes settings. Specifically, our goal is to add novel insights in the following ways: first, we integrate different model-agnostic explanations and provide them to study participants to enable them to assess the decision-making procedures. This contrasts with most existing work, which have typically employed distinct individual explanation styles only. Second, we provide identical model-agnostic explanations to study participants for both the case of ADSs and the human decision maker to not bias the collected responses. Third, we examine how perceptions may change for people with high versus low AI literacy (D. Long & Magerko, 2020). To the best of our knowledge, the combination of the previous aspects has not been examined before. Finally, we consider the provider-customer context of lending, which differentiates our work from M. K. Lee (2018), who has analyzed the perceptions of human versus automated decisions in algorithmic management.

6.3 Research Hypotheses

Drawing on D. Chan (2011, p. 3), *informational fairness* is about "people's expectation that they should receive adequate information on and explanation of the process and its outcomes." In accordance with Bélanger et al. (2002), we define *trustworthiness* as the perception of confidence in the reliability and integrity of the ADS. People often tend to avoid algorithms and prefer a human decision maker over an automated one, even in situations where the algorithm outperforms the person. This phenomenon is called *algorithm aversion* (Dietvorst et al., 2015). Based on this theory, as well as recent developments regarding a decline in trust towards AI (Edelman, 2021), we formulate our first two hypotheses, which conjecture higher perceptions of informational fairness and trustworthiness towards human decision makers as compared to ADSs:

H1 People's perceptions of informational fairness are higher when they are told the decision maker is a human as compared to an ADS.

H2 People's perceptions of trustworthiness are higher when they are told the decision maker is a human as compared to an ADS.

The attitudes towards a decision relevant to their area of expertise may differ significantly between experts of a certain decision-making procedure and laypeople.

For instance, R. Wang et al. (2020) uncovered a notable effect of general computer literacy on fairness evaluations in the context of automated decision-making. In the present study, we assess a construct that is more directly relevant to our context: we gauge individuals' *AI literacy*. This term refers to a set of competencies that empower individuals to critically evaluate AI technologies, interact and collaborate effectively with AI, and utilize AI as a tool online, at home, and in the workplace (D. Long & Magerko, 2020). Our interest lies in determining whether variations in individuals' AI literacy influence their perceptions of informational fairness and trustworthiness regarding human versus automated decision makers. Consequently, we propose the following additional hypotheses:

H3 People's AI literacy moderates the effect of the nature of the decision maker (human versus ADS) on people's perceptions of informational fairness.

H4 People's AI literacy moderates the effect of the nature of the decision maker (human versus ADS) on people's perceptions of trustworthiness.

6.4 Methodology

We evaluate our hypotheses in the context of lending—an example of a provider-customer encounter. Specifically, we confront study participants with situations where a person was denied a loan. We argue that this is a common context that affects many people at some point in life. According to ACTICO (2021), this is also an area where ADSs are commonly employed for high-stakes decision-making.

6.4.1 Study Design

General Setup We choose a between-subjects design with the following conditions: first, we reveal to study participants some basic information about the lending company—similarly to the study setup introduced in some of our earlier work (Schöffer et al., 2021). We then explain that the company rejected a given individual's loan application. Afterwards, we randomly allocate study participants to one of two conditions: 50% of participants are provided the information that an ADS made the decision, and the other 50% are told that the decision maker was a human being. We then provide identical explanations regarding a decision to study participants in either condition, the exact specifications of which will be derived and explained in more detail shortly. Finally, we measure perceptions of informational fairness

(INFF) and trustworthiness (TRST) through multiple measurement items, drawn (and partially adapted) from previous studies (INFF: Colquitt et al. (2001); TRST: Carter and Bélanger (2005), Chiu et al. (2009), and M. K. Lee (2018)). Additionally, we measure AI literacy (AILIT) of study participants, with items partially derived from D. Long and Magerko (2020) as well as Wilkinson et al. (2010). The precise measurement items are identical to the ones that we use in Chapter 5 of this thesis.

Data and ADS We design and implement a functional ADS for our study—similarly to earlier work presented in Chapter 5. The ADS consists of an ML model that predicts loan approval on unseen data and can output different explanations. For training our model, we utilize a publicly available dataset (Chatterjee, 2019) on home loan application decisions. The dataset at hand consists of 614 labeled (loan Y/N) observations. It includes the following features: *applicant income, co-applicant income, credit history, dependents, education, gender, loan amount, loan amount term, marital status, property area, self-employment*. It is worth noting that analogous data, embodying a specific finance company's situation and approval criteria, might practically be employed for training ADSs (Infosys, 2019). After removing data points with missing values, we are left with 480 observations, 332 of which (69.2%) involve the positive label (Y) and 148 (30.8%) the negative label (N). As it is common in ML-based applications, we use 70% of the dataset to train our ADS and use the remaining 30% as a holdout set for the experiment. In preparation for the design of our ADS, we initiate the process by encoding and scaling the features, followed by the training of a random forest classifier (Breiman, 2001). This classifier is capable of predicting unseen, held-out labels with an out-of-bag accuracy of 80.1%, and it serves as the foundation for the scenarios and explanations that the participants encounter.

Explanations Recall that 50% of study participants are assigned the *ADS* condition and 50% the *human* condition. Both conditions are provided with identical explanations regarding the decisions—the only difference is that study participants in the *ADS* condition are told that the ADS provides the explanatory information. In contrast, participants in the human condition are told that a company representative (i.e., a human) provides this information.

We now explain in more detail the provided explanations. As noted earlier, we employ only model-agnostic explanations (Adadi & Berrada, 2018) in a way that they could plausibly be provided by humans and ADSs alike. First, we disclose all *features*, including corresponding values (e.g., *applicant income: $3,069 per month*)

for an observation from the holdout set whom our ADS denied the loan. We refer to such an observation as a *setting*. In our study, we employ different settings to ensure generalizability. We also explain to study participants the *importance* of these features in the decision-making process. For that, we compute permutation feature importances (Breiman, 2001) from our model. For each setting, we finally provide three *counterfactual* scenarios where one actionable feature each is minimally altered such that our model predicts a loan approval instead of a rejection (e.g., *the individual would have been granted the loan if, everything else unchanged, the co-applicant income had been at least $800 per month*). The scenarios that study participants see look similar to the one in Figure 5.2 on page 118 (Chapter 5), except that the introductory text mentions that either an ADS or a human being was the final decision maker.

6.4.2 Data Collection

We conducted a between-subjects online study to test our hypotheses. Participants for this study were recruited via Prolific (Palan & Schitter, 2018) and randomly assigned to either the human decision scenario or the ADS decision scenario. Every participant was provided with two questionnaires associated with two different settings. In each questionnaire, we asked participants to rate their agreement with multiple statements per construct on 5-point Likert scales. A score of 1 corresponds to "strongly disagree" and a score of 5 to "strongly agree." To enrich our understanding of the participants' quantitative responses, we also incorporated several open-ended questions. Out of the 200 responses we gathered, we had to exclude 4 due to their failure to pass an attention check, leaving us with 196 responses for analysis. In our participant pool, males constituted 62%, females 36%, and the remaining 2% either identified as non-binary or chose not to disclose their gender. As for their occupations, 42% were students, 29% were in full-time employment, 11% were part-time workers, 7% were self-employed, 10% were unemployed, and 1% preferred not to reveal their professional status. The average age of the participants stood at 26.4 years.

6.5 Quantitative and Qualitative Results

Before conducting our tests, we assess the validity and reliability of our latent constructs (INFF, TRST, AILIT), each of which is measured through multiple items. We note that the average variance extracted (AVE) is above or equal to the recommended

threshold of 0.5 for INFF and TRST, while the AVE of AILIT is 0.39. According to Fornell and Larcker (1981), if the AVE value of a construct is low, its convergent validity can still be sufficient if composite reliability (CR) is above 0.6. The CR of all our three constructs, INFF (0.83), TRST (0.94), and AILIT (0.72) is, in fact, above the threshold of 0.7, which is recommended by Barclay et al. (1995). Therefore, our convergent validity is sufficient for AILIT as well. Values for Cronbach's alpha (CA) are larger than the recommended threshold of 0.7 for our three constructs, indicating good reliability (Cortina, 1993). Validity and reliability measures are summarized in Table 6.1.

Tab. 6.1.: Correlations and measurement information for latent constructs.

Construct	M	SD	CA	CR	AVE	INFF	TRST	AILIT
INFF	3.57	0.62	0.83	0.83	0.50	1.00		
TRST	3.45	0.72	0.94	0.94	0.72	0.69	1.00	
AILIT	2.87	0.61	0.71	0.72	0.39	0.30	0.27	1.00

Note: M = Mean; SD = Standard deviation; CA = Cronbach's alpha; CR = Composite reliability; AVE = Average variance extracted.

6.5.1 Comparison of Perceptions

We conduct two Mann-Whitney U tests (McKnight & Najab, 2010) to examine the differences in perceptions between ADS and human decision makers. The Mann-Whitney U test for informational fairness is statistically significant $(p = 0.017)$, suggesting a significant difference between participants' perceptions of informational fairness. Comparing the means of perceptions of informational fairness for both conditions reveals that the ADS condition $(M = 3.68)$ is perceived to be significantly fairer than the human condition $(M = 3.47)$. For perceptions of trustworthiness, however, there is no significant difference between the conditions $(p = 0.113)$. Hence, neither **H1** nor **H2** are supported by our analyses. In fact, **H1** is reversely supported, eventually suggesting that for our study setup, informational fairness perceptions tend to be *higher towards the ADS* compared to the human decision maker. Based on qualitative responses from study participants, we conjecture that this might be due to the perceived absence of emotions and subjectivity in automation. Other potential reasons for this based on qualitative feedback are given in Section 6.5.2. Note that this finding seems contradictory to some prior works' results (e.g., Castelo et al. (2019)), which raises doubts about the generalizability of such findings beyond specific domains.

Interestingly, when considering people's AI literacy, these results change. For this analysis, we split our data into two (approximately equal-sized) sub-samples along the median value of AI literacy. We refer to one sample as *high AI literacy* participants and the other as *low AI literacy* participants. We then conduct separate Mann-Whitney U tests for the two sub-samples. Participants with high AI literacy perceive the ADS as significantly more informationally fair ($p = 0.021$) and more trustworthy ($p = 0.042$) than the human decision maker. For participants with low AI literacy, we do not find a significant difference for perceptions of informational fairness ($p = 0.312$) or trustworthiness ($p = 0.995$) between the human and the ADS condition. Hence, we conclude that AI literacy has a moderating effect, which supports **H3** and **H4**. As stated in Section 6.3, we expected the moderating effect of AI literacy. However, the finding that people with high AI literacy tend to perceive ADSs as both fairer and more trustworthy than human decision makers is not obvious to us. On the one hand, we might think that people with high AI literacy understand such systems better and are thus less skeptical; on the other hand, it might well be the case that the same type of people are more aware of the shortcomings of ADSs.

6.5.2 Qualitative Insights

We also collected unstructured textual data based on open-ended questions embedded in our questionnaires. An in-depth analysis reveals that many study participants are convinced that automation is precisely the reason why decisions are fair (*"Automated system is fair by design"*). They perceive the ADS as fair because, in their opinion, its decisions are objective: *"It [the ADS] states the criteria and follows [them], there is no room for subjectivity and the data is used to make an objective decision."* This is likely one of the reasons why our hypotheses **H1** and **H2** are not supported. While some participants allude to underlying issues of automated decisions (*"AI can be programmed to be unfair"* and *"I do not believe an Automated Decision System can replace a human. We can't expect it to not make mistakes"*), most view the ADS as fair because the system is *"purely looking at numbers [therefore] its* [sic] *completely fair."* Finally, one person points out that the situation *"is fair because the consumer knows that he has been judged using an algorithm."*

On the other hand, an interesting comment states that *"[t]he decision may have been made by a machine, but someone decided to program it that way,"* which raises questions around accountability of ADSs. Some issues are equally criticized in the human and the ADS condition: *"I don't think it is fair to take education, gender or marital status into account,"* or *"[s]ome factors are indifferent to the decision of the loan and are personal information."* Even though overall the human condition is

perceived as significantly less informationally fair than the ADS condition and people believe the ADS *"can help eliminate [...] bias,"* there are still participants who *"hope bots wont* [sic] *have to decide crucial life decisions for [them]."*

6.6 Conclusion and Outlook

We conducted an online study with 200 participants to evaluate differences in people's perceptions of informational fairness and trustworthiness towards human versus automated decision-making in the high-stakes context of lending. We provided thorough explanations to study participants, identical in both conditions (human and automated), to facilitate meaningful and unbiased responses. Our findings suggest that within the scope of our study setup—contrary to some prior work as well as our own hypothesis—automated decisions are perceived as more informationally fair than human-made decisions. In contrast, no significant differences were measured for trustworthiness in our case. Based on qualitative responses, it appears that people particularly appreciate the absence of subjectivity in ADSs as well as their data-driven approach. Interestingly, our analyses also imply that people's AI literacy affects their perceptions, given the provided explanations. Specifically, we found that people with high AI literacy tend to perceive ADSs as both fairer and more trustworthy than a human decision maker, whereas no significant differences for either construct were detected for people with low AI literacy.

Based on our findings, we may conjecture that providing thorough explanations can enhance perceptions of fairness and trustworthiness towards ADSs over human decision makers—particularly for people with higher AI literacy. However, we must be cognizant of the dangers of wrongful persuasion and automation biases, that is, the tendency of people to over-rely on ADSs (Goddard et al., 2012; Skitka et al., 2000). This might become a problem if too many (compelling) explanations about the inner workings of ADSs are provided. Future work should also account for this by examining how perceptions change when the quality of the ADS changes for the worse (e.g., by making unfair decisions), similar to the blueprint suggested by Schöffer and Kühl (2021). Other natural extensions include the consideration of domains other than lending, as well as the adoption of different explanation styles. We hope that our work will stimulate multifaceted future research on this topic of utmost societal relevance.

Part III

Human-In-The-Loop Decision-Making

7 Conceptualizing the Interdependence of Reliance Behavior and Decision Quality

In this chapter, we establish a formal relationship between reliance behavior and decision-making accuracy in the context of human-in-the-loop decision-making. Crucially, we demonstrate that the capacity of humans to complement an artificial intelligence (AI) system hinges on three factors: (i) the baseline AI accuracy, (ii) the quantity of human reliance, and (iii) the quality of their reliance. Given these three dimensions, we propose a visual framework intended to serve as a blueprint for interpreting and comparing empirical results in human-in-the-loop decision-making. From this framework, we derive several interesting properties that emphasize the importance of concurrently measuring reliance behavior *and* accuracy when empirically assessing the impacts of transparency mechanisms or other interventions.

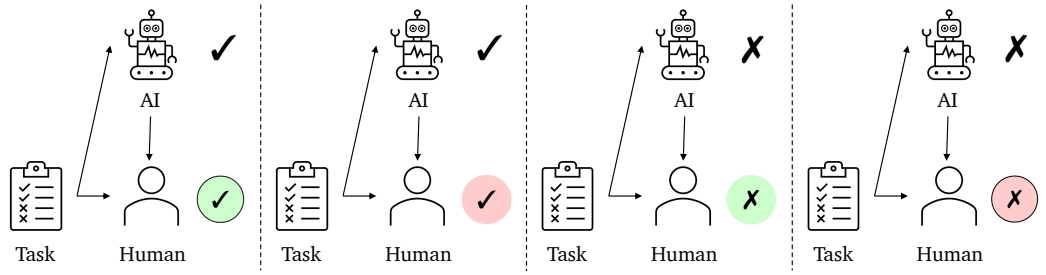
7.1 Introduction

Decision-making increasingly leverages support from AI systems with the goal of making better and more efficient decisions. Especially in high-stakes domains, such as lending, hiring, or healthcare, researchers and policymakers have often advocated for having a human-in-the-loop as the "last line of defense against AI failures" (Passi & Vorvoreanu, 2022, p. 1). This assumes that humans can correct such AI failures in

This chapter is based on published work. To enhance the reading experience and maintain overall consistency of the thesis, we removed the abstract and made several minor adjustments. The original paper can be accessed via:

Schöffer, J.,* Jakubik, J.,* Vössing, M., Kühl, N. & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. *HHAI 2023: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence* (pp. 46–59). https://doi.org/10.3233/FAIA230074.
*denotes equal contribution

Fig. 7.1.: Different types of human reliance on AI recommendations.



(a) Correct adherence (b) Wrong overriding (c) Correct overriding (d) Wrong adherence

Note: We consider concurrent *decision-making setups where a human decision maker receives a task and corresponding AI recommendation that can either be correct (✓) or wrong (✗), as indicated by the respective symbol next to the AI. The human can then either adhere to (bordered circle) or override (no border) the AI recommendation. When the human adheres to correct or overrides a wrong AI recommendation, the final decision will be correct (cases (a) and (c)); in the remaining cases, it will be wrong (cases (b) and (d)). The correctness of the final decision is indicated by either green or red shading.*

the first place. In human-in-the-loop settings, typically, an AI system generates an initial decision recommendation, which the human may either adhere to or override (see Figure 7.1). In the taxonomy of Tejeda et al. (2022), this corresponds to *concurrent AI assistance*, where the human does *not* independently make a decision before AI assistance is provided. In order to complement the AI system, the human would have to adhere to its recommendations if and only if these recommendations are correct, and override them otherwise. Empirical studies have shown, however, that humans are often not able to achieve this type of *appropriate reliance* (Bansal et al., 2021; Passi & Vorvoreanu, 2022; Schemmer et al., 2023). Instead, we often observe that humans either over- or under-rely on AI recommendations, or simply cannot calibrate their reliance.[1] Even the introduction of additional means of decision support (e.g., explanations) has often not shown the expected benefits in terms of enabling humans to complement AI systems. Worryingly, any root cause analyses are hindered by the fact that the mechanisms through which such interventions affect humans' reliance behavior are not well understood.

In this work, we make explicit and analyze the interplay of human reliance behavior on AI recommendations and decision-making accuracy, and we highlight the importance of assessing and reporting *both* in empirical studies on human-in-the-loop decision-making. To this end, we develop a framework that disentangles reliance *quantity* and *quality*, and lets us understand how both—individually and

[1] We use *reliance* as an umbrella term for humans' behavior of adhering to or overriding AI recommendations (Lai et al., 2021).

in conjunction—affect decision-making accuracy. We also visualize these interdependencies geometrically, which aims at making them easier to grasp. The visual framework is ultimately intended to serve researchers for interpreting empirical findings, including the effects of interventions, in human-in-the-loop decision-making. It may also be used by practitioners to reflect on their reliance behavior when interacting with AI systems.

From our theoretical analyses, we infer several interesting properties: *first*, we show that over- and under-reliance are not symmetric with respect to their effects on decision-making accuracy. Specifically, when humans adhere too little to recommendations from an AI system that performs better than chance, it is impossible to improve decision-making accuracy over the AI baseline. *Second*, when humans are unable to distinguish correct from wrong AI recommendations, that is, when their reliance behavior is independent of the correctness of AI recommendations, we cannot expect humans to complement an AI system, either. In such cases, we also see that "blindly" adhering more to AI recommendations increases the expected decision-making accuracy—without any improved ability to discern correct and wrong recommendations. Finally, *third*, we show that interventions may affect accuracy through drastically different effects on reliance. For instance, two different interventions may lead to an identical increase in accuracy, but one may do so through *decreasing* human adherence to AI recommendations, whereas the other may lead to an *increase* in adherence. Both interventions may look identically effective when not analyzing effects on reliance behavior at the level that we propose. These insights are crucial for designing meaningful decision support measures.

7.2 Background

Measuring and calibrating the human reliance on AI recommendations has become a central pillar of research on AI-informed decision-making (Lai et al., 2021; Passi & Vorvoreanu, 2022; Schemmer et al., 2023). This is especially important as both humans and AI systems are imperfect "decision makers" with individual strengths and weaknesses (Bansal et al., 2021; Hemmer et al., 2021; Mozannar & Sontag, 2020). For humans that are assisted by AI systems, it is therefore essential to be able to identify strengths and weaknesses of the AI system; that is, in which cases it is correct and in which wrong (Schemmer, Hemmer, Kühl, et al., 2022). In this setting, latest research distinguishes three cases of reliance behavior: (i) relying on AI recommendations in too few cases (i.e., *under-reliance* (Lu & Yin, 2021; Schuetz et al., 2022), e.g., by underestimating AI performance), (ii) relying on AI recommendations

in too many cases (i.e., *over-reliance* (Buçinca et al., 2021; Passi & Vorvoreanu, 2022; Vasconcelos et al., 2023), e.g., by overestimating AI performance), and (iii) relying *appropriately* on AI recommendations (i.e., adhering to AI recommendations when correct and overriding when wrong (Ashktorab et al., 2021; Schemmer, Hemmer, Kühl, et al., 2022)). Thus far, research has identified many scenarios in which under-reliance or over-reliance results in reduced decision-making performance (Buçinca et al., 2021; Kim et al., 2023). In an emerging effort, more works are developed around achieving an appropriate level of reliance, which is a prerequisite for the human decision maker to complement the AI system and ultimately improve the overall decision-making accuracy over the AI baseline (Schemmer et al., 2023). In this work, we develop a framework that aims at improving our understanding of how human reliance behavior translates to decision-making accuracy.

Accuracy of AI-informed decision-making (i.e., the number of correct decisions given the overall number of decisions) represents a key metric that may indicate the utility of an AI system—apart from other metrics such as fairness (De-Arteaga et al., 2022). The accuracy metric is hence frequently used for measuring the performance of human-in-the-loop decision-making (Lai et al., 2021) and evaluating the effectiveness of interventions (e.g., explanations) for decision support (Cabrera et al., 2023; Lai et al., 2020, 2021; Y. Zhang et al., 2020). Overall, we observe that research has typically focused on either the performance in terms of accuracy (Kim et al., 2023; Lai et al., 2021) or on the human behavior in terms of their reliance on AI recommendations (Buçinca et al., 2021; Lu & Yin, 2021), when assessing effects of interventions. However, in human-in-the-loop decision-making, accuracy is immediately influenced by the degree to which humans adhere to or override AI recommendations, and *how* they do so (Jakubik et al., 2023). In this work, we show that the relationship between reliance behavior and accuracy follows clear mathematical patterns, and that measuring either decision-making accuracy or the level of reliance alone may provide an incomplete view when assessing human-in-the-loop decision-making generally and the effects of interventions specifically.

7.3 The Interdependence of Reliance Behavior and Accuracy

For clarity of exposure, we consider binary decision-making tasks of $n \in \mathbf{N}$ instances with n AI recommendations. Let $Acc_{AI} \in (50\%, 100\%]$ be the AI accuracy,[2] and $\mathcal{A} \in$

[2]Note that we only consider cases where the AI performs strictly better than chance.

Tab. 7.1.: We distinguish four cases of human reliance behavior in binary human-in-the-loop decision-making.

	Correct AI	**Wrong AI**
Adherence to AI	Correct adherence ($\mathcal{A}_{correct}$)	Wrong adherence ($\mathcal{A}_{wrong}$)
Overriding of AI	Wrong override ($\mathcal{O}_{wrong}$)	Correct override ($\mathcal{O}_{correct}$)

$[0\%, 100\%]$ the degree of human adherence to AI recommendations—for instance, $\mathcal{A} = 70\%$ when the human adheres to 70% of AI recommendations. As introduced in Figure 7.1 on page 150, adherence can be correct ($\mathcal{A}_{correct}$) or wrong ($\mathcal{A}_{wrong}$), and we have $\mathcal{A} = \mathcal{A}_{correct} + \mathcal{A}_{wrong}$. Similarly, we call the number of overrides $\mathcal{O} \in [0\%, 100\%]$ (correct: $\mathcal{O}_{correct}$ or wrong: $\mathcal{O}_{wrong}$), and we have $\mathcal{O} = \mathcal{O}_{correct} + \mathcal{O}_{wrong}$. While in practice humans can only adhere to or override an integer number of AI recommendations, we often consider $n \to \infty$ for our theoretical considerations, so as to avoid rounding. We summarize the possible cases of adhering and overriding AI recommendations in Table 7.1. Note that by definition we also have:

$$\begin{aligned} \mathcal{A} + \mathcal{O} &= \mathcal{A}_{correct} + \mathcal{A}_{wrong} + \mathcal{O}_{correct} + \mathcal{O}_{wrong} = 100\% \\ Acc_{AI} &= \mathcal{A}_{correct} + \mathcal{O}_{wrong} \\ Acc_{final} &= \mathcal{A}_{correct} + \mathcal{O}_{correct}. \end{aligned} \tag{7.1}$$

7.3.1 Motivational Example

Consider the following motivational example: we have a task that consists of making $n = 10$ binary decisions. The AI system that is used for providing decision recommendations to the human has an accuracy of $Acc_{AI} = 70\%$; that is, 7 out of 10 recommendations are correct (✓) and 3 are wrong (✗). Now, when the human adheres to all AI recommendations ($\mathcal{A} = 100\%$), this leads to a decision-making accuracy of $Acc_{final} = 70\%$, equal to the AI accuracy. In terms of reliance behavior, this implies that the human correctly adhered to 7 correct AI recommendations ($\mathcal{A}_{correct} = 70\%$), and wrongly adhered to the remaining 3 recommendations ($\mathcal{A}_{wrong} = 30\%$). In the other extreme case where the human overrides all AI recommendations ($\mathcal{O} = 100\%$), the resulting decision-making accuracy will be $100\% - 70\% = 30\%$, where the human correctly overrides 3 wrong AI recommendations ($\mathcal{O}_{correct} = 30\%$), and wrongly overrides 7 correct AI recommendations ($\mathcal{O}_{wrong} = 70\%$).

If the human reliance behavior is mixed, that is, when the human adheres to some AI recommendations and overrides others, decision-making accuracy will depend on how well the human can distinguish cases where the AI is correct from cases

Fig. 7.2.: Possible scenarios of reliance behavior and associated decision-making accuracy, given an AI accuracy of $Acc_{AI} = 70\%$ and an adherence level of $\mathcal{A} = 70\%$.

Note: Correct AI recommendations (✓) and wrong AI recommendations (✗) are separated by a dashed line.

where it is wrong. To make this clear, consider the same AI as above with an accuracy of 70%, and a human that adheres to 7 out of 10 of its recommendations ($\mathcal{A} = 70\%$). This is illustrated in Figure 7.2. If the human is able to perfectly distinguish between correct and wrong AI recommendations, they will adhere to all 7 correct AI recommendations ($\mathcal{A}_{correct} = 70\% = \mathcal{A}$) and override the 3 wrong ones ($\mathcal{O}_{correct} = 30\% = \mathcal{O}$). The resulting decision-making accuracy would then be $Acc_{final} = 100\%$ (case (a) in Figure 7.2). In this case, the human is able to perfectly complement the AI system by correcting for its mistakes. Cases (b)–(d) in Figure 7.2 show situations where the human still adheres to 70% of AI recommendations but their ability to override wrong AI recommendations decreases. For instance, consider case (d), where the human does not perform any correct overrides ($\mathcal{O}_{correct} = 0$). When the human degree of adherence to AI recommendations is fixed at 70% this is, in fact, the worst possible reliance behavior with respect to accuracy, resulting in a decision-making accuracy of $Acc_{final} = 40\%$.

From Figure 7.2 on page 154, we can also infer that if the human overrides *more* than 3 AI recommendations, at least one of these overrides must be wrong (i.e., the human would override a correct AI recommendation), meaning that a decision-making accuracy of 100% would no longer be possible. We may think of such a reliance behavior as *under-reliance*. Similarly, when the human overrides *less* than 3 AI recommendations, there must be at least one instance of wrong adherence. This might be referred to as *over-reliance*. In the general case, we may think of under-reliance as a behavior where $\mathcal{A} < Acc_{AI}$, and over-reliance as $\mathcal{A} > Acc_{AI}$. Note that there exists other work that has been thinking of these terms with respect to behavior at the level of individual decisions (Schemmer et al., 2023; Vasconcelos et al., 2023).

7.3.2 The General Case

Generally, any degree of adherence to AI recommendations is associated with a range of possible decision-making accuracy, based on how well the human can override the AI recommendations when they are wrong and adhere to them when they are correct. In Figure 7.2 on page 154, this range would be $Acc_{final} \in \{40\%, 60\%, 80\%, 100\%\}$ for $n = 10$, a given AI accuracy of $Acc_{AI} = 70\%$, and a degree of adherence to AI recommendations of $\mathcal{A} = 70\%$. As mentioned earlier, we generally consider $n \to \infty$, in which case this range becomes continuous. We state the following proposition[3] on the attainable decision-making accuracy as a function of the AI accuracy as well as the degree of human adherence to AI recommendations.

Proposition 7.1. *For $n \to \infty$, a given AI accuracy Acc_{AI}, and a degree of adherence to AI recommendations $\mathcal{A}$, the range of attainable decision-making accuracy Acc_{final} is*

$$Acc_{final} \in \begin{cases} [100\% - Acc_{AI} - \mathcal{A}, 100\% - Acc_{AI} + \mathcal{A}] & \text{if } 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ [-100\% + Acc_{AI} + \mathcal{A}, 100\% - Acc_{AI} + \mathcal{A}] & \text{if } 100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI} \\ [-100\% + Acc_{AI} + \mathcal{A}, 100\% + Acc_{AI} - \mathcal{A}] & \text{if } Acc_{AI} < \mathcal{A} \leq 100\%. \end{cases}$$

The maximum of this accuracy range will be attained whenever the human maximizes correct adherence and correct overrides given a degree of adherence $\mathcal{A}$, since $Acc_{final} = \mathcal{A}_{correct} + \mathcal{O}_{correct}$. Hence, in the ideal case, we would have $\mathcal{A}_{correct} + \mathcal{O}_{correct} = 100\%$; which immediately implies that $\mathcal{A}_{wrong} = \mathcal{O}_{wrong} = 0\%$.

[3] All of our theoretical results can be readily derived from the visual framework that is introduced in Section 7.3.3. Hence, we refrain from providing formal proofs in this chapter.

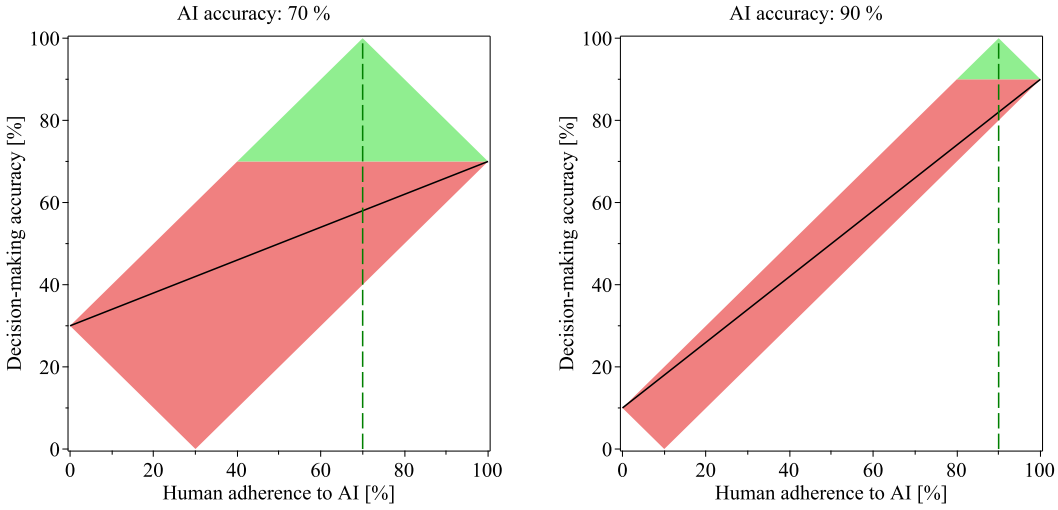
This would be case (a) in Figure 7.2 on page 154. However, as we can see in Proposition 7.1, this is only possible when $\mathcal{A} = \mathcal{A}_{correct} = Acc_{AI}$, meaning that the human must adhere to AI recommendations if and only if they are correct, and override otherwise. In other words, to achieve a decision-making accuracy of $Acc_{final} = 100\%$, we need two things:

(i) The human's general degree of adherence to AI recommendations, $\mathcal{A}$, is equal to the AI accuracy Acc_{AI}, that is, $\mathcal{A} = Acc_{AI}$.

(ii) The human must be able to adhere to any correct AI recommendation and override any wrong one, that is, $\mathcal{A}_{correct} = \mathcal{A}$ and $\mathcal{O}_{correct} = \mathcal{O}$.

However, in practice, it is likely that either (i) or (ii) are not satisfied and, hence, the decision-making accuracy is less than 100%. Even if (i) is satisfied, like in Figure 7.2 on page 154, we see in cases (b)–(d) that Acc_{final} is negatively affected when humans adhere to wrong AI recommendations and override correct ones.

Fig. 7.3.: Visual framework on the interdependence of reliance behavior and decision quality.



(a) $Acc_{AI} = 70\%$ (b) $Acc_{AI} = 90\%$

Note: The area of attainable decision-making accuracy for a given AI accuracy is colored. The red area indicates $Acc_{final} < Acc_{AI}$; green indicates $Acc_{final} > Acc_{AI}$; the green dashed line indicates the level of adherence where $Acc_{final} = 100\%$ is attainable; the black line indicates the expected Acc_{final} when humans cannot discern correct and wrong AI recommendations.

7.3.3 A Visual Framework

To make the general relationship between reliance behavior and decision-making accuracy more tangible, we visualize Proposition 7.1 in Figure 7.3 on page 156 for (a) $Acc_{AI} = 70\%$ and (b) $Acc_{AI} = 90\%$. On the horizontal axes we have the human adherence to AI recommendations, $\mathcal{A} \in [0, 100\%]$. The vertical axes show the decision-making accuracy, $Acc_{final} \in [0, 100\%]$. The filled rectangular area in red and green combined constitutes the attainable decision-making accuracy for any given $\mathcal{A}$. We distinguish red and green to highlight areas where the human-in-the-loop complements the AI (green, $Acc_{final} > Acc_{AI}$) or impairs it (red, $Acc_{final} < Acc_{AI}$) regarding accuracy. The green dashed vertical line indicates the level of $\mathcal{A} = Acc_{AI}$, which corresponds to the degree of adherence where the maximum decision-making accuracy of 100% can be attained, as discussed previously. Note that as the AI accuracy increases ((a) $\rightarrow$ (b) in Figure 7.3 on page 156), the colored area decreases; and for $Acc_{AI} = 100\%$ it becomes a line, in which case Proposition 7.1 collapses into $Acc_{final} = \mathcal{A}$.

Contrasting the red and green areas, we immediately see that up to a certain level of $\mathcal{A}$ there is no possibility to reach the green area, where $Acc_{final} > Acc_{AI}$. We also see that the minimum level of $\mathcal{A}$ for which the human-in-the-loop may complement the AI increases as Acc_{AI} increases ($\mathcal{A} = 40\% \rightarrow \mathcal{A} = 80\%$ in Figure 7.3 (a) $\rightarrow$ (b) on page 156). Finally, when $\mathcal{A} \geq Acc_{AI}$, attaining a decision-making accuracy in the green area is always possible. We characterize this in the following corollary:

Corollary 7.1. *When humans under-rely at a degree of $\mathcal{A} < 2 \cdot Acc_{AI} - 100\%$, we will always have $Acc_{final} < Acc_{AI}$. When $\mathcal{A} > 2 \cdot Acc_{AI} - 100\%$, achieving a decision-making accuracy greater than the AI accuracy, that is, $Acc_{final} > Acc_{AI}$, is possible.*

From the visual framework, we also see that any $Acc_{final} \in (0, 100\%)$ can be associated with different degrees of adherence $\mathcal{A}$. In fact, due to the symmetric shape of the rectangle, when we think of Acc_{final} as a function of $\mathcal{A}$, the inverse $\mathcal{A}(Acc_{final})$ would be identical to the function itself. For instance, a decision-making accuracy of $Acc_{final} = 70\%$ may correspond to any $\mathcal{A} \in [40\%, 100\%]$.

Proposition 7.2. *When $Acc_{final}(\mathcal{A}) \in [u, v]$ for a given $\mathcal{A}$, we have $\mathcal{A}(Acc_{final}) = [u, v]$.*

However, fixing Acc_{final} at 70%, different levels of $\mathcal{A}$ correspond to different vertical positions within the rectangle: $\mathcal{A} = 40\%$ corresponds to a position at the very northern border of the rectangle, whereas any $\mathcal{A} \in [70\%, 100\%]$ corresponds to a position on the horizontal line that separates the red and green areas. This means that a given decision-making accuracy can be achieved through strikingly different reliance behaviors. We address this, as well as the role of the black separating lines in Figure 7.3 on page 156, in more detail in the following.

7.3.4 Discerning Correct and Wrong AI Recommendations

While a horizontal movement in the framework constitutes a change in the *quantity* of adherence to AI recommendations, this information alone does not capture the *quality* of reliance. This is captured in the vertical movements. To make this more concrete, consider again a task with AI recommendations that are 70% accurate. When the human has no ability to distinguish correct from wrong AI recommendations, the likelihood of adhering to or overriding a given AI recommendation is the same regardless of whether that recommendation is correct or wrong. Hence, at an adherence of $\mathcal{A}$, we would expect the human to adhere to $\mathcal{A}\%$ of correct AI recommendations and $\mathcal{A}\%$ of wrong AI recommendations. At $Acc_{AI} = 70\%$, this implies that $\mathcal{A}\%$ of 70% are correct adherences, $\mathcal{A}\%$ of 30% are wrong adherences, $(100 - \mathcal{A})\%$ of 70% are wrong overrides, and $(100 - \mathcal{A})\%$ of 30% are correct overrides. When we have $\mathcal{A} = 70\%$, this would imply $\mathcal{A}_{correct} = 49\%$, $\mathcal{A}_{wrong} = 21\%$, $\mathcal{O}_{correct} = 9\%$, and $\mathcal{O}_{wrong} = 21\%$, with a decision-making accuracy of $\mathcal{A}_{correct} + \mathcal{O}_{correct} = 58\%$. This corresponds to the intersection of the black line with the dashed green vertical line in Figure 7.3 (a) on page 156. We generalize this in the following proposition.

Proposition 7.3. *When humans cannot discern correct and wrong AI recommendations, the expected decision-making accuracy is linearly increasing in $\mathcal{A}$ and given by*

$$\begin{aligned} Acc_{final}(\mathcal{A}) &= \mathcal{A} \cdot Acc_{AI} + (100\% - \mathcal{A}) \cdot (100\% - Acc_{AI}) \\ &= (100\% - Acc_{AI}) + \underbrace{(2 \cdot Acc_{AI} - 100\%)}_{>0} \cdot \mathcal{A}, \end{aligned}$$

for a given AI accuracy Acc_{AI}.

Note that the relationship from Proposition 7.3 equates to the black lines in Figure 7.3 on page 156, which separate the respective rectangles in half. We immediately see the following:

Corollary 7.2. *When humans cannot discern correct and wrong AI recommendations, the expected decision-making accuracy is always lower or equal to the AI accuracy, that is,* $Acc_{final} \leq Acc_{AI}$.

Having established the expected decision-making accuracy when humans are not able to distinguish correct and wrong AI recommendations, we now turn to cases where they can—to different degrees. Such reliance behavior corresponds to points in the framework that are situated *above* the black line. While certainly less relevant in practice, we might also think of cases where humans adhere to and override AI recommendations worse than chance, which would correspond to points *below* the black line. Following up on Proposition 7.1, we now examine three cases based on different levels of adherence to AI recommendations, and we characterize the reliance behavior that is associated with the maximum and minimum decision-making accuracy for given $\mathcal{A}$.

Case 1 We first consider $0 \leq \mathcal{A} \leq 100\% - Acc_{AI}$. Since we assume that $Acc_{AI} > 50\%$, we have $\mathcal{A} < Acc_{AI}$ in this case. When the degree of adherence to AI recommendations is strictly smaller than the AI accuracy, achieving a decision-making accuracy of $Acc_{final} = 100\%$ is no longer possible. This also implies that there must be at least one instance where the human overrides a correct AI recommendation, that is, $\mathcal{O}_{wrong} > 0$. From Proposition 7.1 we also see that the *maximum* achievable decision-making accuracy in that case is $100\% - Acc_{AI} + \mathcal{A}$, which is achieved when $\mathcal{A}_{correct} = \mathcal{A}$. Using the definition of $\mathcal{A}$ and relationships from Equation (7.1), this directly implies that $\mathcal{A}_{wrong} = 0$, $\mathcal{O}_{correct} = 100\% - Acc_{AI}$, and $\mathcal{O}_{wrong} = Acc_{AI} - \mathcal{A} > 0$. The *minimum* achievable decision-making accuracy, on the other hand, is attained when adherence only happens to wrong AI recommendations, hence, $\mathcal{A}_{wrong} = \mathcal{A}$. Similar to above, we this implies that $\mathcal{A}_{correct} = 0$, $\mathcal{O}_{wrong} = Acc_{AI}$, and $\mathcal{O}_{correct} = 100\% - Acc_{AI} - \mathcal{A}$.

To illustrate this, let us reconsider the example from Figure 7.2 on page 154, but with a degree of adherence to AI recommendations of $\mathcal{A} = 20\%$. The attainable decision-making accuracy in this case is, according to Proposition 7.1, $Acc_{final} \in [10\%, 50\%]$. To achieve $Acc_{final} = 50\%$, the human would have to adhere to 2 correct AI recommendations ($\mathcal{A}_{correct} = 20\%$) and 0 wrong AI recommendations ($\mathcal{A}_{wrong} = 0$).

The remaining 8 AI recommendations, 5 of which are correct and 3 wrong, are overridden (i.e., $\mathcal{O}_{wrong} = 50\%$ and $\mathcal{O}_{correct} = 30\%$). The minimum decision-making accuracy of 10%, on the other hand, is attained when the human only adheres to wrong AI recommendations (i.e., $\mathcal{A}_{wrong} = 20\%$ and $\mathcal{A}_{correct} = 0$). The remaining AI recommendations, 7 correct and 1 wrong, are overridden, which implies $\mathcal{O}_{wrong} = 70\%$ and $\mathcal{O}_{correct} = 10\%$. Overall, we conclude the following:

Corollary 7.3. *When $0 \leq \mathcal{A} \leq 100\% - Acc_{AI}$, the decision-making accuracy is maximal when all adherence is to correct AI recommendations (i.e., $\mathcal{A}_{correct} = \mathcal{A}$), and it is minimal when all adherence is to wrong AI recommendations (i.e., $\mathcal{A}_{wrong} = \mathcal{A}$).*

Case 2 Let us now consider $100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI}$. With the same argument as in the previous case, the *maximum* decision-making accuracy is attained when $\mathcal{A}_{correct} = \mathcal{A}$, which directly implies $\mathcal{A}_{wrong} = 0$, $\mathcal{O}_{correct} = 100\% - Acc_{AI}$, and $\mathcal{O}_{wrong} = Acc_{AI} - \mathcal{A}$. As for the *minimum* decision-making accuracy, note that since $\mathcal{A} > 100\% - Acc_{AI}$, we must have $\mathcal{A}_{correct} > 0$, that is, the human must be adhering to at least one correct AI recommendation. The minimum accuracy is thus attained when the human adheres to all wrong AI recommendations plus at least one correct recommendation. This implies that all overrides must be of correct AI recommendations, that is, we have $\mathcal{O}_{wrong} = \mathcal{O}$, $\mathcal{O}_{correct} = 0$, as well as $\mathcal{A}_{correct} = Acc_{AI} - \mathcal{O} > 0$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI}$.

Corollary 7.4. *When $100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI}$, the decision-making accuracy is maximal when all adherence is to correct AI recommendations (i.e., $\mathcal{A}_{correct} = \mathcal{A}$), and it is minimal when all overrides are of correct AI recommendations (i.e., $\mathcal{O}_{wrong} = \mathcal{O}$).*

Case 3 Finally, we consider cases where $Acc_{AI} < \mathcal{A} \leq 100\%$. While in the previous two cases we had $\mathcal{A} \leq Acc_{AI}$, we now consider the case where humans over-rely on the AI recommendations, meaning that there must be a least one case where the human adheres to a wrong AI recommendation, that is, $\mathcal{A}_{wrong} > 0$. The *maximum* decision-making accuracy will thus be attained when all overrides are correct, that is, $\mathcal{O}_{correct} = \mathcal{O}$, which immediately implies $\mathcal{O}_{wrong} = 0$, $\mathcal{A}_{correct} = Acc_{AI}$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI} - \mathcal{O} > 0$. The *minimum* decision-making accuracy, on the other hand, will be attained when all overrides are wrong, similar

to the previous case. Hence, we would also observe $\mathcal{O}_{wrong} = \mathcal{O}$, $\mathcal{O}_{correct} = 0$, $\mathcal{A}_{correct} = Acc_{AI} - \mathcal{O} > 0$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI}$.

Corollary 7.5. *When $Acc_{AI} < \mathcal{A} \leq 100\%$, the decision-making accuracy is maximal when all overrides are of wrong AI recommendations (i.e., $\mathcal{O}_{correct} = \mathcal{O}$), and it is minimal when all overrides are of correct AI recommendations (i.e., $\mathcal{O}_{wrong} = \mathcal{O}$).*

7.3.5 Measuring the Quality of Reliance

In the previous subsection, we established the reliance behavior that is associated with the extreme cases of maximum and minimum decision-making accuracy for any given degree of adherence to AI recommendations. Now, we develop a metric $Q(\mathcal{A}) \in [0, 1]$ for the quality of reliance given Acc_{AI}, such that a value of 1 corresponds to the maximum attainable decision-making accuracy, and 0 to the minimum. First, we derive the following corollary from Proposition 7.1:

Corollary 7.6. *The width W of the range of attainable values for Acc_{final} is:*

$$W = \begin{cases} 2 \cdot \mathcal{A} & \text{if } \ 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ 2 \cdot (100\% - Acc_{AI}) & \text{if } \ 100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI} \\ 2 \cdot (100\% - \mathcal{A}) & \text{if } \ Acc_{AI} < \mathcal{A} \leq 100\%. \end{cases}$$

Geometrically, W corresponds to the distance between the upper and lower vertical boundary of the rectangle (see Figure 7.3 on page 156) for a fixed $\mathcal{A}$. With that, we can define our metric $Q(\mathcal{A})$ as follows:

$$Q(\mathcal{A}) \coloneqq \begin{cases} \dfrac{(\mathcal{A}_{correct} + \mathcal{O}_{correct}) - (100\% - Acc_{AI} - \mathcal{A})}{W} & \text{if } \ 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ \dfrac{(\mathcal{A}_{correct} + \mathcal{O}_{correct}) + (100\% - Acc_{AI} - \mathcal{A})}{W} & \text{if } \ 100\% - Acc_{AI} < \mathcal{A}. \end{cases} \tag{7.2}$$

Note that since Acc_{AI} and $\mathcal{A}$ are fixed, maximizing the quality of reliance, $Q(\mathcal{A})$, corresponds to maximizing $\mathcal{A}_{correct} + \mathcal{O}_{correct} = Acc_{final}$, and we have seen what this entails in terms of reliance behavior for any value of $\mathcal{A}$ in Section 7.3.4. Note that $Q(\mathcal{A})$ is not constant in cases where humans cannot discern correct and wrong

AI recommendations. In this case, using Proposition 7.3, we obtain $Q(30\%) = 0.7$, whereas $Q(70\%) = 0.3$. We may think of this as follows: while $\mathcal{A} = 70\%$ leads to a higher expected decision-making accuracy of $Acc_{final} = 58\%$ (compared to $Acc_{final} = 42\%$ for $\mathcal{A} = 30\%$), the attainable accuracy in either case is $[40\%, 100\%]$ at $\mathcal{A} = 70\%$, and $[0, 60\%]$ in the case of $\mathcal{A} = 30\%$. Hence, the accuracy relative to the "potential" is much worse in the case of $\mathcal{A} = 70\%$. This will be relevant in the following section.

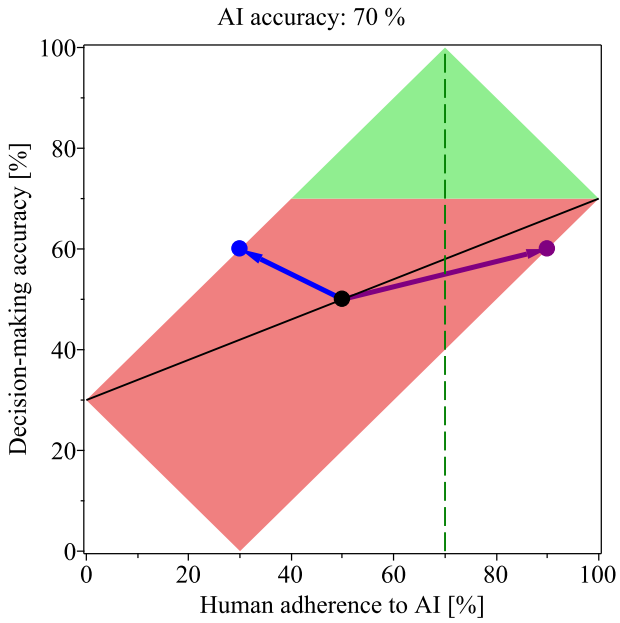
7.4 Understanding the Effects of Interventions

Our visual framework can be used to depict empirical results in human-in-the-loop decision-making and understand them better. Any such empirical finding would be a static point in the colored rectangle, from which we can immediately infer interesting properties, such as the quantity and quality of reliance, the exact percentages of correct adherence and overrides, or the ability of the human to complement or not the AI.

Another key usage of the framework is its ability to understand and disentangle the effects of interventions, such as explanations or other means of decision support (Lai et al., 2021). For that, let us consider the following hypothetical example: through a randomized experiment, we have collected data where humans are making decisions in the presence of two different types of explanations (• and •) versus a baseline without explanations (•). We can think of these interventions as movements in our visual framework, as depicted in Figure 7.4 on page 163. The black point corresponds to a situation where a human cannot discern correct and wrong AI recommendations and adheres to $\mathcal{A} = 50\%$.

Now, in the case of the blue intervention, we see that it leads to a decrease in the degree of adherence to AI recommendations, compared to the baseline ($\mathcal{A} = 50\% \rightarrow \mathcal{A} = 30\%$), but an increase in decision-making accuracy ($Acc_{final} = 50\% \rightarrow Acc_{final} = 60\%$) through a better reliance quality ($Q = 0.5 \rightarrow Q = 1$). In the case of the purple intervention, we see the same effect with respect to accuracy but an entirely different effect on the reliance behavior—where this intervention leads to an *increase* in adherence to AI recommendations ($\mathcal{A} = 50 \rightarrow \mathcal{A} = 90\%$). At the same time, reliance quality drops from $Q = 0.5$ to $Q = 0$, which from Corollary 7.5 we know corresponds to a situation of over-reliance where any of the $\mathcal{O} = 10\%$ overrides are of correct AI recommendations. Finally, note that since the purple point lies below the black line, this corresponds to reliance behavior that is of lower quality

Fig. 7.4.: Visualizing the effects of different interventions (• and •) on reliance behavior and decision-making accuracy.



according to Equation (7.2) than in cases where the human decides at random which AI recommendations to adhere to or override. This implies that different interventions can have seemingly similar effects on decision-making accuracy but drastically different effects on reliance behavior. Our framework enables us to disentangle these dimensions.

7.5 Discussion and Conclusion

In this work, we propose a framework to understand and analyze the interdependence between reliance behavior and decision-making accuracy in human-in-the-loop decision-making. We show that any given *quantity* of humans' adherence to AI recommendations is associated with a specific range of attainable decision-making accuracy, depending on the *quality* of reliance, that is, humans' ability to adhere to AI recommendations if and only if they are correct. Vice versa, we also show that any accuracy level can be achieved through fundamentally different reliance behavior, both in terms of reliance quantity and quality. This has implications for assessing the effectiveness of interventions, such as explanations or other forms of decision support, in human-in-the-loop decision-making. For instance, our work highlights the importance of assessing and reporting *both* effects on accuracy *and* reliance

behavior in order to derive meaningful implications on how such interventions affect decision-making. Specifically, we show an example of how assessing only effects on accuracy may lead to the wrong conclusion that an intervention was not effective when in reality it changed reliance behavior significantly. Even more worryingly, by not measuring or reporting effects on reliance behavior, we may conclude that an intervention led to an increase in decision-making accuracy, without understanding that this increase was driven solely by an increase in adherence *quantity* while the ability to discern correct and wrong AI recommendations dropped.

We also infer interesting properties when the human-in-the-loop cannot discern correct and wrong AI recommendations, that is, when the probability of adhering to or overriding a given AI recommendation is independent of its correctness. In practice, this may occur when a task is too difficult for the human to solve. In such cases, we show that the human may never be expected to complement the AI system, meaning that the decision-making accuracy will be strictly lower than the initial AI accuracy—except when the human adheres to *all* AI recommendations, in which case the decision-making accuracy will be equal to the AI accuracy. Another interesting implication of this analysis is that expected decision-making accuracy is linearly increasing in the quantity of adherence to AI recommendations, that is, decision-making accuracy may be increased by solely adhering to more AI recommendations. This must be considered when interpreting empirical findings.

Finally, we infer that under- and over-reliance[4] is not symmetrical regarding their implications for decision-making accuracy. While the human may complement the AI system when over-relying by systematically adhering to correct recommendations and overriding wrong ones, there is no hope for improvements in accuracy over the AI baseline when the human under-relies past a threshold of $\mathcal{A} < 2 \cdot Acc_{AI} - 100\%$. Notably, this threshold may be very high when the AI system performs well—for instance, at an AI accuracy of 90%, any adherence $\mathcal{A} < 80\%$ can *never* lead to a decision-making accuracy that is better than the AI system. Especially when the human-in-the-loop is not aware of such high AI performance, it might be unrealistic to expect them to complement the AI system.

Our framework is currently applicable to binary decision-making tasks with an AI system in place that performs better than chance. A natural extension would be to include cases with more than two decision alternatives. In such cases, our reliance taxonomy would have to be altered to account for situations where overriding a wrong AI recommendation may still lead to a wrong decision. Our visual framework is also limited in its use for situations where we want to compare empirical findings

[4]Recall that we define *under-reliance* globally as $\mathcal{A} < Acc_{AI}$, and vice versa for over-reliance.

across studies with *different* AI accuracy. Extending it to account for varying AI accuracy would involve a 3-dimensional visual with a third axis on Acc_{AI}. Finally, we might think of cases where the metric of decision-making performance is not accuracy but, for instance, fairness. In these instances, our framework can offer valuable insights when used to evaluate the impacts of interventions on different types of prediction errors, specifically when these errors are distinguished based on a sensitive attribute like gender or race.

8 Assessing Transparency Effects on Reliance Behavior and Fairness of Decisions

In this chapter, we empirically investigate the link between feature-based explanations, distributive fairness, and human reliance on recommendations from artificial intelligence (AI) systems. Our results indicate that explanations shape fairness perceptions, influencing humans' propensity to follow AI recommendations. However, these explanations do not reliably enable humans to distinguish correct from wrong—instead, they impact reliance regardless of AI recommendation accuracy. Depending on the highlighted features, we also see that explanations can either promote or impede distributive fairness. When explanations highlight task-irrelevant features that are associated with gender, they prompt anti-stereotypical overrides of AI recommendations. Conversely, task-relevant explanations can intensify stereotype-aligned errors by the AI system. These findings suggest that feature-based explanations are not a reliable mechanism for enhancing distributive fairness, as their effectiveness hinges on the flawed idea of "fairness through unawareness."

8.1 Introduction

AI systems are commonly used for informing decision-making in consequential areas, where they provide human decision makers with decision recommendations. The human is then tasked to decide whether to adhere to these recommendations or override them. Researchers, policymakers, and activists have expressed concern over the risk of algorithmic bias resulting in unfair decisions. As a response, many

This chapter is based on work that is currently under review as follows:

> Schöffer, J., De-Arteaga, M.* & Kühl, N.* (2023). On explanations, fairness, and appropriate reliance in human-AI decision-making. *Under Review. Preliminary Version: ACM CHI 2023 Workshop on Trust and Reliance in AI-Assisted Tasks (TRAIT)*.
> *denotes equal contribution

have advocated for the need for explanations, under the assumption that they can enable humans to mitigate algorithmic bias. For instance, in a recent Forbes article (Kite-Powell, 2022, p. 1), it is claimed that "companies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums." Others have claimed that explanations "provide a more effective interface for the human in-the-loop, enabling people to identify and address fairness and other issues" (Dodge et al., 2019, p. 275). However, there is often ambiguity regarding what it means for the human to mitigate bias, and a lack of evidence studying whether this is possible. In this work, we posit that when concerned with distributive fairness, the central mechanism that should be studied is the type of reliance fostered by the explanations and its effect on disparities in AI-informed decisions.

Our Work In this work, we examine the effects of feature-based explanations on people's ability to enhance distributive fairness—and how these effects are mediated by fairness perceptions and reliance on AI recommendations. To empirically study this, we conduct a randomized online experiment and assess differences in perceptions and reliance behavior when participants see and do not see explanations, and when these explanations indicate the use of sensitive features in predictions versus when they indicate the use of task-relevant features. We operationalize this study in the context of occupation prediction, for which we train two AI systems with access to different vocabularies. We randomly assign participants to one of two groups and ask them to predict whether bios belong to professors or teachers: for one group, recommendations come from a model that uses *gendered* words for predicting occupations, whereas in the other group the model uses *task-relevant* words. Both models provide the same recommendations, and their distribution of errors is in line with societal stereotypes and the expected risks of bias characterized in previous research (De-Arteaga et al., 2019). Participants in both conditions are provided with explanations that visually highlight the most predictive words of their respective models. We also include a baseline condition where no explanations are shown. We test for differences in perceptions and reliance behavior across conditions, and measure gender disparities for different types of errors.

Findings and Implications *First*, we do not observe any significant differences in decision-making accuracy across conditions, that is, participants did not make more (or less) accurate decisions in the conditions with explanations compared to the baseline without explanations. Since participants were incentivized to make

accurate predictions, this implies that explanations did not enable them to make better decisions with respect to accuracy.

Second, no condition improved participants' likelihood to override mistaken versus correct AI recommendations, but conditions did affect the likelihood to override AI recommendations conditioned on the predicted occupation: we see that participants in the *gendered* condition overrode more AI recommendations to *counter* existing societal stereotypes (e.g., by predicting more women to be professors), irrespective of whether the prediction was correct. Simultaneously, when explanations highlight only task-relevant words, reliance behavior *reinforced* stereotype-aligned decisions; for instance, by predicting more men to be professors, even when they are teachers.

This, *third*, has implications for distributive fairness: by prompting reliance behavior that either counters or reinforces societal stereotypes embedded in AI recommendations, (i) explanations that highlight gendered words led to a *decrease* in error rate disparities (i.e., fostering distributive fairness), whereas (ii) explanations that highlight task-relevant words led to an *increase* in error rate disparities (i.e., hindering distributive fairness). These findings emphasize the need to differentiate between improved distributive fairness that is driven by a shift in the types of errors versus improvements that are driven by humans' ability to override mistaken AI recommendations.

Fourth, we confirm prior works' findings by observing that people's fairness perceptions are significantly lower when explanations highlight gendered words compared to task-relevant words, and empirically show that people override significantly more AI recommendations when their fairness perceptions are low. However, we observe that perceptions solely relate to the quantity of overrides and do *not* correlate with an ability to discern correct and wrong AI recommendations. Hence, fairness perceptions are only a meaningful proxy for distributive fairness when it is desirable to override the AI system based on its use of sensitive features. However, prior research has shown that the idea of "fairness through unawareness" (Kusner et al., 2017, p. 2) is neither a necessary nor sufficient condition for distributive fairness (Apfelbaum et al., 2010; Corbett-Davies & Goel, 2018; Dwork et al., 2012; Kleinberg et al., 2018; Nyarko et al., 2021; Pedreshi et al., 2008).

8.2 Background

In this section, we provide background on our work and review related literature on explanations, reliance, and fairness.

8.2.1 Explanations of AI

Goals of Explanations AI systems are becoming increasingly complex and opaque, and researchers and policymakers have called for explanations to make AI systems more understandable to humans (European Union, 2016; Langer, Oster, et al., 2021; J. Miller & Chamberlin, 2000). Apart from the central aim of facilitating human understanding, prior research has formulated a wealth of different desiderata that explanations are to provide, most of which center one or more different types of stakeholders of AI systems (Ehsan & Riedl, 2020; Langer, Oster, et al., 2021; Preece et al., 2018). For instance, system developers might be interested in facilitating trust in their systems through explanations, whereas a regulator likely wants to assess a system's compliance with moral and ethical standards (Langer, Oster, et al., 2021). Different goals may sometimes be impossible to accomplish simultaneously (Springer & Whittaker, 2019). Relevant to our work are several desiderata that concern explanations as an alleged means for better and fairer AI-informed decision-making (Adadi & Berrada, 2018; Dodge et al., 2019); we speak to this in more detail in Sections 8.2.2 and 8.2.3. For a comprehensive overview of different aims of explanations, we refer the reader to Langer, Oster, et al. (2021) and Lipton (2018).

Types of Explanations The scientific literature distinguishes explanations that aim at explaining individual predictions (*local* explanations) from those that aim at explaining the general functioning of an AI system (*global* explanations), as summarized by Guidotti et al. (2018). However, it has been argued that combining local explanations can also lead to an understanding of global model behavior (Lundberg et al., 2020). So-called *local model-agnostic* explanations, such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017), have gained popularity in the literature (Adadi & Berrada, 2018). When applied to text data, these methods can generate a highlighting of important words for text classification. In this work, our focus is on these feature-based explanations, and we use LIME in our experiments, due to its popularity in the literature as well as in practice (Bhatt et al., 2020; ElShawi et al., 2021; Gilpin et al., 2018).

Criticism of Explanations Most desiderata for explanations are insufficiently studied or met with inconclusive or seemingly contradictory empirical findings (V. Chen et al., 2023; de Bruijn et al., 2022; Langer, Oster, et al., 2021). A major line of criticism stems from the fact that explanations can mislead people: Chromik et al. (2019) discuss situations where system developers may create interfaces or

misleading explanations to purposefully deceive more vulnerable stakeholders like auditors or decision subjects; for instance, through *adversarial attacks* on explanation methods (Dimanov et al., 2020; Lakkaraju & Bastani, 2020; Pruthi et al., 2020; Slack, Hilgard, et al., 2020). In the extreme case of placebic explanations (i.e., explanations that convey no information about the underlying AI), Eiband et al. (2019, p. 1) find that people may exhibit levels of trust similar to "real explanations." This shows that the sheer presence of explanations can increase people's trust in AI. Even in the absence of any malicious intents, Ehsan and Riedl (2021) highlight several challenges arising from unanticipated negative downstream effects of explanations, such as misplaced trust in AI, or over- or underestimating the AI's capabilities. In the context of fairness, feature-based explanations may or may not highlight the usage of sensitive information (e.g., on gender) by an AI system, which has been shown to be an unreliable indicator of a system's actual fairness (Apfelbaum et al., 2010; Corbett-Davies & Goel, 2018; Dwork et al., 2012; Kleinberg et al., 2018; Nyarko et al., 2021; Pedreshi et al., 2008). We address this in more detail in Section 8.2.3 due to its importance for our work.

8.2.2 Explanations and (Appropriate) Reliance

Effects on Accuracy It has been argued that explanations are an enabler for better human-in-the-loop decision-making (Arrieta et al., 2020; Dodge et al., 2019; Gilpin et al., 2018; Kizilcec, 2016; Rader et al., 2018). A recent meta-study (Schemmer, Hemmer, Nitsche, et al., 2022) on the effectiveness of explanations, however, implies that explanations in most empirical studies did not yield any significant benefits with respect to decision-making accuracy; for instance, in the studies of Alufaisan, Marusich, et al. (2021), Green and Chen (2019b), H. Liu et al. (2021), M. Narayanan et al. (2018), and Y. Zhang et al. (2020). On the other hand, Lai and Tan (2019) find that explanations greatly enhance decision-making accuracy for the case of deception detection. An accuracy increase through explanations may, however, solely be due to (i) an overall increase in adherence to a high-accuracy AI system, or (ii) an overall decrease in adherence to a low-accuracy AI system (Schöffer, Jakubik, et al., 2023).

Effects on Reliance In the context of human-in-the-loop decision-making, *appropriate reliance* is typically understood as the behavior of humans of overriding wrong AI recommendations and adhering to correct ones (Passi & Vorvoreanu, 2022; Schemmer, Hemmer, Kühl, et al., 2022). Humans' ability to override mistaken

recommendations has also been referred to as *corrective overriding* (De-Arteaga et al., 2020). When considering the role of explanations in fostering appropriate reliance, it has been claimed that "transparency mechanisms also function to help users to learn about how the system works, so they can evaluate the *correctness* of the outputs they experience and identify outputs that are incorrect" (Rader et al., 2018, p. 3). Empirical evidence, however, is less clear: several studies have found that explanations can be detrimental to appropriate reliance, when they increase or decrease humans' adherence to AI recommendations regardless of their correctness (Bansal et al., 2021; Bussone et al., 2015; Lai et al., 2021; Poursabzi-Sangdeh et al., 2021; Schemmer, Kühl, Benz, & Satzger, 2022; van der Waa et al., 2021). These phenomena are commonly referred to as *over-* or *under-reliance* (Schemmer, Hemmer, Kühl, et al., 2022).

Conflation of Reliance and Trust Many studies have treated reliance and trust interchangeably (Lai et al., 2021), sometimes calling reliance a "behavioural trust measure" (Papenmeier et al., 2022, p. 18). However, definitions of *trust* are often inconsistent (Jacovi et al., 2021; J. D. Lee & See, 2004; Papenmeier et al., 2022), which makes empirical findings challenging to compare. More importantly, trust and reliance are different constructs (Lai et al., 2021): reliance is the *behavior* of adhering to or overriding AI recommendations, whereas trust is a subjective *attitude* regarding the whole system, which builds up and develops over time (Parasuraman & Riley, 1997; Rempel et al., 1985; K. Yu et al., 2017). It has been argued that trust may impact reliance (Dzindolet et al., 2003; J. D. Lee & See, 2004; Shin & Park, 2019), but trust is not a sufficient requirement for reliance when other factors, such as time constraints, perceived risk, or self-confidence, impact decision-making (De-Arteaga et al., 2020; J. D. Lee & See, 2004; Riley, 2018). In our work, we directly measure participants' reliance behavior and do not assume an equivalence between reliance and trust.

8.2.3 Explanations and Fairness

Goal of Promoting Algorithmic Fairness It is known that AI systems can issue predictions that may result in disparate outcomes or other forms of injustices for certain socio-demographic groups—especially those that have been historically marginalized (Bartlett et al., 2022; Buyl et al., 2022; De-Arteaga et al., 2022; Imana et al., 2021). When AI systems are used to inform consequential decisions, it is important that a human can override problematic recommendations. To that end, the literature

has often framed explanations as an important pathway towards improving algorithmic fairness (Arrieta et al., 2020; Das & Rad, 2020; Dodge et al., 2019; Langer, Oster, et al., 2021). Grounded on the organizational justice literature (Colquitt & Rodell, 2015; Greenberg, 1987), researchers distinguish different notions of algorithmic fairness, among which are (i) *distributive fairness*, which refers to the fairness of decision outcomes (Zafar et al., 2019), and (ii) *procedural fairness*, which refers to the fairness of decision-making procedures (M. K. Lee et al., 2019). Distributive fairness is typically measured in terms of statistical metrics such as parity in error rates across groups (Barocas et al., 2019; Chouldechova, 2017); which is closely related to notions like *equalized odds* or *equal opportunity* (Hardt et al., 2016). Importantly, there is no conclusive evidence showing that explanations lead to fairer decisions, and it remains unclear *how* explanations may enable this (Langer, Oster, et al., 2021).

Fairness Perceptions Prior work at the intersection of fairness and explanations has primarily focused on assessing how people *perceive* the fairness of AI systems (Lai et al., 2021; Starke et al., 2022). Empirical findings are mostly inconclusive, stressing that fairness perceptions depend on many factors, such as the explanation style (Binns et al., 2018; Dodge et al., 2019), the amount of information provided (Schöffer, Kühl, & Machowski, 2022), the use case (Angerschmid et al., 2022), user profiles (Dodge et al., 2019), or the decision outcome (Shulner-Tal et al., 2022). Surprisingly, only few works have examined downstream effects of fairness perceptions on AI-informed decisions. Our work complements prior studies by centering distributive fairness and how it relates to fairness perceptions.

Perceptions and Sensitive Features A series of prior studies have found that knowledge about the features that an AI system uses influences people's fairness perceptions (Grgić-Hlača, Redmiles, et al., 2018; Grgić-Hlača, Zafar, et al., 2018; Grgić-Hlača et al., 2020; Nyarko et al., 2021; Plane et al., 2017; van Berkel et al., 2019). This type of information is, for instance, conveyed by feature-based explanations like LIME. Specifically, people tend to be averse to the use of what is typically considered *sensitive* information, for instance, gender or race (Corbett-Davies & Goel, 2018; Grgić-Hlača, Redmiles, et al., 2018; Grgić-Hlača, Zafar, et al., 2018; Grgić-Hlača et al., 2020; Nyarko et al., 2021; Plane et al., 2017; Schöffer, Kühl, & Machowski, 2022). Interestingly, people's perceptions towards these features change after they learn that "blinding" the AI system to these features can lead to *worse* outcomes for marginalized groups. Similarly, it has been shown that people's perceptions towards the inclusion of sensitive features switch when they are told that this inclusion makes

an AI system more accurate (Grgić-Hlača et al., 2020) or equalizes error rates across demographic groups (Harrison et al., 2020). In fact, it is known that prohibiting an AI system from using sensitive information is neither a necessary nor sufficient requirement for fair decision-making (Apfelbaum et al., 2010; Corbett-Davies & Goel, 2018; Dwork et al., 2012; Kleinberg et al., 2018; Nyarko et al., 2021; Pedreshi et al., 2008), and that there exist several real-world examples where the inclusion of sensitive features can make historically disadvantaged groups like Black people or women better off (Corbett-Davies & Goel, 2018; Mayson, 2019; Pierson et al., 2020; Skeem et al., 2016). In this work, we build upon these findings on the interplay of fairness perceptions and sensitive features. Concretely, we assess differences in reliance behavior when participants see explanations that highlight task-relevant versus sensitive features, and derive implications for distributive fairness.

8.3 Study Design

In this section, we outline our study design. First, we introduce the task and dataset for our study, then we explain the experimental setup and our dependent variables, and, finally, the data collection process.

8.3.1 Task and Dataset

Task Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online (Bogen & Rieke, 2018; Sánchez-Monedero et al., 2020). An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI systems, it is susceptible to gender bias and discrimination (Bogen & Rieke, 2018; De-Arteaga et al., 2019; Sánchez-Monedero et al., 2020). De-Arteaga et al. (2019) show that these biases can manifest themselves in error rate disparities between genders, and that error rate disparities are correlated with gender imbalances in occupations. For instance, women surgeons are significantly more often misclassified than men surgeons because the occupation *surgeon* is heavily men-dominated. Similar disparities occur, among others, for professors and teachers. Interestingly, the disparate impact on people persists when the AI system does *not* consider explicit gender indicators, such as pronouns (De-Arteaga et al., 2019).

Such misclassifications in hiring have tremendous repercussions for affected people because they may be systematically excluded from exposure to relevant opportunities. In our study, we instantiate a human-in-the-loop decision-making setup where participants see short textual bios and are asked—with the help of an AI recommendation—to predict whether a given bio belongs to a professor or a teacher. Professors are historically a men-dominated occupation, whereas teachers have been mostly associated with women (J. Miller & Chamberlin, 2000).[1]

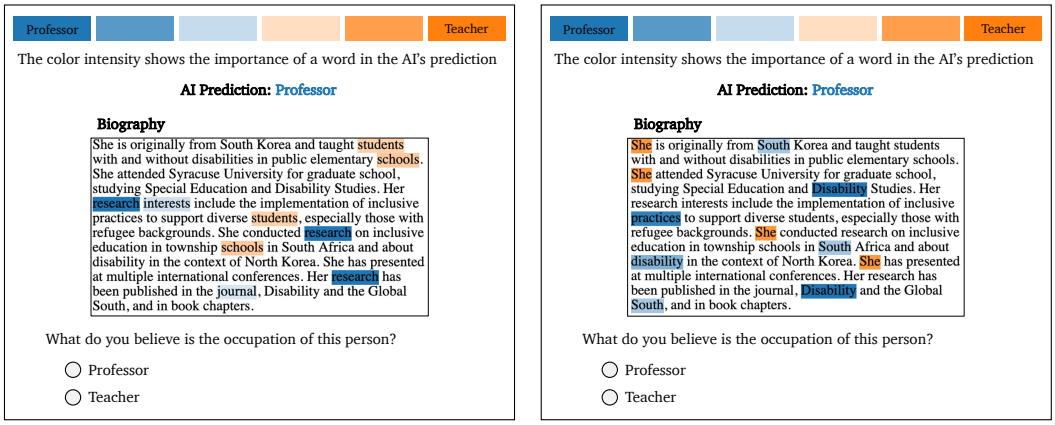
Dataset We use the publicly available BIOS dataset, which contains approximately 400,000 online bios for 28 different occupations from the Common Crawl corpus, initially created by De-Arteaga et al. (2019). This dataset has been used in other studies on AI-informed decision-making as well, such as the ones by H. Liu et al. (2021) and Peng et al. (2022). For each bio in the dataset we know the gender of the corresponding person and their true occupation. Gender is based on the pronouns used in the bio, and a limitation of this dataset is that it only contains bios that use "she" or "he" as pronouns, excluding bios of non-binary people (Gorny et al., 2023a, 2023b). We only consider bios that belong to professors and teachers, which leaves us with 134,436 bios, out of which 118,215 belong to professors and 16,221 to teachers. In line with current demographics and societal stereotypes (J. Miller & Chamberlin, 2000; Zippia, 2022a, 2022b), we have more men (55%) than women (45%) bios of professors, and more women (60%) than men (40%) bios of teachers.

8.3.2 Experimental Setup

General Procedure Participants see 14 bios one by one, each including the AI recommendation as well as an explanation highlighting the most predictive words. We also include a baseline condition without explanations. The crux of our experimental design is that we assign participants to conditions where they see recommendations and explanations either from (i) an AI system that uses *task-relevant* features, or (ii) an AI system that uses *gendered* (i.e., sensitive) features. An exemplary bio including explanations is depicted in Figure 8.1 on page 176. Note that the AI predictions and explanations stem from actual AI systems that agree in their predictions for the 14 bios shown to participants; we outline the construction of these models later in Section 8.3.3.

[1] See also Zippia (2022a, 2022b) on current demographic statistics for professors and teachers in the US.

Fig. 8.1.: A bio of a woman professor, in the (a) *task-relevant* and the (b) *gendered* condition.



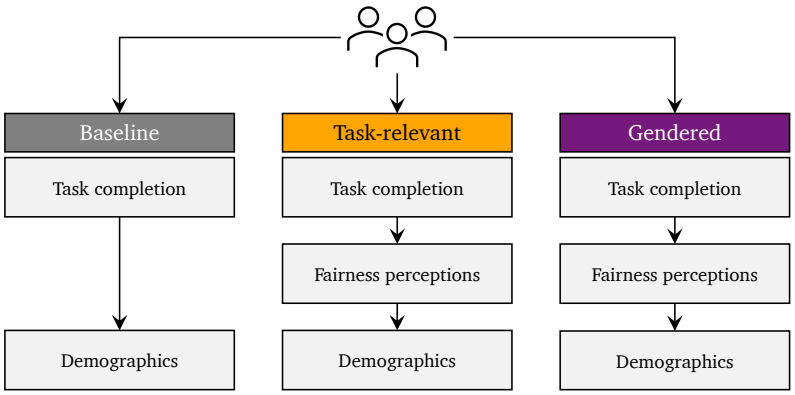
(a) *Task-relevant* condition

(b) *Gendered* condition

Participants in each condition first complete the task of predicting occupations for 14 bios, and—if assigned to a condition with explanations—answer several questions regarding their fairness perceptions. Since the baseline condition does not provide any cues regarding the AI system's decision-making procedures, we do not ask about perceptions there. Finally, participants provide some demographic information. A summary of our general setup in illustrated in Figure 8.2 on page 177. Note that we ask about fairness perceptions *after* the task is completed, so as to prevent these questions from moderating reliance behavior (Chaudoin et al., 2021). Given that distinguishing professors and teachers based on their bios can be at times ambiguous and not everyone may be familiar with the differences, we also ask at the beginning of our questionnaires what participants consider the difference between *professor* and *teacher* to be. Additionally, after completing the task, we ask participants an open-ended question on what information they relied on when differentiating *professor* and *teacher*. This way, we were able to confirm—both quantitatively and qualitatively—that participants thought consistently about this distinction between conditions.

Task Completion Figure 8.1 shows the interface that participants in the *task-relevant* as well as the *gendered* condition see during the completion of the task. Explanations involve a dynamic highlighting of important words for either AI system (*task-relevant* and *gendered*); and they also indicate whether certain words are indicative of *professor* (blue) or *teacher* (orange). Lastly, the color intensity shows the importance of a given word in the AI prediction. This interface is similar to

Fig. 8.2.: Illustration of our experimental setup.



Note: Study participants are randomly assigned to one of three conditions. In each condition, they first complete the task of predicting occupations from 14 short bios, and complete a demographic survey. In the conditions with explanations, participants are also asked about their fairness perceptions after completing the task. We use this color-coding (grey/orange/purple) throughout the thesis to refer to our conditions.

related studies on AI-assisted text classification (Lai et al., 2020; H. Liu et al., 2021; Schemmer, Hemmer, Kühl, et al., 2022). Participants in the *task-relevant* and the *gendered* condition are confronted with 14 bios similar to the one in Figure 8.1 on page 176, whereas participants in the baseline condition are shown the same set of bios without highlighting of words, and the AI prediction without color coding. Recall that the AI recommendations are identical across conditions. For each instance, participants are asked to make a binary prediction about whether they believe that a given bio belongs to a professor or a teacher. We incentivize accurate predictions through bonus payments (see Section 8.3.6).

8.3.3 Task-Relevant and Gendered Classifiers

We now explain in more detail how we constructed the AI systems that we use for generating recommendations and explanations in the *task-relevant* and *gendered* conditions. The general idea is to train two classifiers with access to mutually disjoint feature sets (i.e., vocabularies). The *task-relevant* vocabulary consists of words that appear on average—for both men and women—more often in professor or teacher bios than in any of the 26 remaining occupations in the BIOS dataset. The *gendered* vocabulary, on the other hand, consists of words that are most predictive of gender.

Let $\mathcal{W} := \{w_1, \ldots, w_n\}$ be the set of n words that occur most often across the set of all bios. We chose $n = 5000$, that is, $\mathcal{W}$ contains the top-5000 most occurring words,

after removal of (manually defined) stop words. We inferred $\mathcal{W}$ from applying a `CountVectorizer` (Pedregosa et al., 2011). In trial runs, we found that increasing n further does not significantly change the classifiers' predictions. We then constructed two logistic regression classifiers, $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$, with access to mutually disjoint vocabularies: *task-relevant words* ($\mathcal{W}_{rel} \subset \mathcal{W}$) and *gendered words* ($\mathcal{W}_{gen} \subset \mathcal{W}$).

Task-Relevant Vocabulary We performed the following steps to construct the task-relevant vocabulary $\mathcal{W}_{rel}$:

1. For all $i \in \{1, \ldots, n\}$, compute the average occurrence of word $w_i \in \mathcal{W}$ in bios of men and women professors and teachers. We call the results $\widehat{w_i^{P,m}}$, $\widehat{w_i^{P,w}}$, $\widehat{w_i^{T,m}}$, and $\widehat{w_i^{T,w}}$, where we use P, T and m, w as a shorthand for the respective occupations and genders. We also compute $\widehat{w_i^{\bullet}}$ as the average occurrence of w_i for any other occupation $\bullet$ that is *not* professor or teacher.
2. For given gender $g \in \{m, w\}$, check if $\widehat{w_i^{P,g}} > \widehat{w_i^{\bullet}}$ *or* $\widehat{w_i^{T,g}} > \widehat{w_i^{\bullet}}$ for all other occupations $\bullet$, that is, whether the average occurrence of word w_i in professor or teacher bios of gender g is greater than the average in *any* other occupation. If this condition is met, add w_i to $\mathcal{W}_{rel}^{g}$, the set of task-relevant words for gender g.
3. Compute $\mathcal{W}_{rel}^{m} \cap \mathcal{W}_{rel}^{w} = \mathcal{W}_{rel}$ as the set of words that are task-relevant for *both* genders.

After completing steps 1–3, we obtain the task-relevant vocabulary $\mathcal{W}_{rel}$ of 543 words, including *faculty*, *kindergarten*, or *phd*, among others.

Gendered Vocabulary Denote $|\mathcal{B}^{o,g}|$ the amount of bios of occupation $o \in \{P, T\}$ and gender $g \in \{m, w\}$. We perform the following steps to construct the gendered vocabulary $\mathcal{W}_{gen}$:

1. Sample equal amounts of bios for men and women professors and teachers. Since $\min\{|\mathcal{B}^{o,g}|\} = |\mathcal{B}^{T,m}| = 6440$, randomly sample 6440 bios for each combination of occupation and gender.
2. Extract features from bios by applying a `CountVectorizer` with `TF-IDF` weighting (Pedregosa et al., 2011).
3. Train a logistic regression to predict *gender* from the extracted features.

4. Compute the importance of each (weighted) feature based on the absolute magnitude of their corresponding regression coefficient, and sort the resulting list of words by importance.

5. Include the top-5% most important words in $\mathcal{W}_{gen}$ as the set of words that are highly predictive of gender. We choose the threshold of 5% so as to exclude words that are spuriously correlated with gender (e.g., *towards*).

After completing steps 1–5, we obtain the gendered vocabulary $\mathcal{W}_{gen}$ of 214 words, which include—apart from gender pronouns and words such as *husband* and *wife*—words like *dance*, *art*, or *engineering*, which are not evidently gendered but highly correlated with the sensitive attribute.

Deploying the Classifiers Having established our vocabularies $\mathcal{W}_{rel}$ and $\mathcal{W}_{gen}$, we proceed by training two logistic regression[2] models on a balanced set of bios containing 50% professors and 50% teachers. Denote $|\mathcal{B}^P|$ and $|\mathcal{B}^T|$ the amounts of bios of occupations P and T. Since $|\mathcal{B}^T| = 16,221 < |\mathcal{B}^P|$, we randomly sample 16,221 bios of professors, while preserving the gender distribution from the original data. This yields a dataset of 32,442 bios, 50% of which we use as a holdout set. We separate a relatively large holdout set because we will eventually use a specific subset of these bios in our questionnaires (see Section 8.3.4). The resulting classifiers achieve F_1 scores of 0.87 ($\mathbf{AI}_{rel}$) and 0.77 ($\mathbf{AI}_{gen}$). For generating dynamic explanations with highlighting of predictive words, we employ the `TextExplainer` from LIME (Ribeiro et al., 2016).

8.3.4 Selection of Bios

In order to be able to assess differences in reliance behavior across conditions, participants see a mix of cases where the AI recommendations are correct or wrong. More specifically, we distinguish six types of scenarios that make up the 14 bios that participants see—they are summarized in Table 8.1 on page 180. We distinguish these scenarios based on three dimensions: (i) gender of the person associated with a bio; (ii) true occupation of that person; (iii) AI recommended occupation. We show 3 cases each of correctly recommended women teachers (WTT) and men professors (MPP), as well as 3 cases of wrongly recommended women professors (WPT) and men teachers (MTP). Note that our focus is on scenarios where the AI recommendations are in line with gender stereotypes. To preempt the misconception

[2]We use logistic regression to ensure that explanations are faithful to the underlying model.

Tab. 8.1.: Overview of the six types of scenarios employed in our study.

Gender of bio	True occupation	AI recommendation	AI correct?	Acronym	#Bios
Woman	**T**eacher	**T**eacher	✓	WTT	3
Woman	**P**rofessor	**T**eacher	✗	WPT	3
Woman	**P**rofessor	**P**rofessor	✓	WPP	1
Man	**T**eacher	**T**eacher	✓	MTT	1
Man	**T**eacher	**P**rofessor	✗	MTP	3
Man	**P**rofessor	**P**rofessor	✓	MPP	3

that the AI system always recommends *teacher* for women and *professor* for men, we also include one case each of correctly recommended woman professor (WPP) and correctly recommended man teacher (MTT). In the light of recent findings from Kim et al. (2023), we include the WPP and MTT scenarios early on in our questionnaires. Precisely, we randomize the order in which participants see the 14 bios, with the restriction that the WPP and MTT scenarios are shown among the first five. We do not consider scenarios where women teachers are classified as professors, or where men professors are classified as teachers, because our focus is on the errors that are more likely to occur in practice (De-Arteaga et al., 2019).

Screening As outlined in Section 8.3.2, participants are confronted with 14 bios of professors and teachers. All bios shown to participants are taken from a random holdout set of BIOS that our two classifiers make predictions on. We impose a series of constraints to select which bios from the holdout set we include in the questionnaires. In particular, for a given bio to be included in our questionnaires, we require it to satisfy the following:

- Both models $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ must yield the same predicted occupation for the bio.
- The prediction probabilities of $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ towards either occupation must be *at most* 20% different. This ensures that both models are comparably certain in their predictions for the given bio.
- The prediction probabilities of $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ towards either occupation must be *at most* 80%. This aims at eliminating a large share of bios that are too easy to classify.
- To avoid any confounding effects of bios' length on people's behavior, we only consider bios of length between 50 and 100 words.

Enforcing these constraints on bios from the holdout set leaves us with 690 eligible bios (out of 16,221). In a next step, we decide on the final set for our questionnaires.

Final Selection The authors jointly screened these 690 bios and ruled out those that are trivial (e.g., because humans would easily be able to tell the occupation) or otherwise not suitable (e.g., because of misspellings or excessive use of jargon). We also discarded bios where explanations would highlight too few or too many words, or where the number of highlighted words was significantly different between the *task-relevant* and the *gendered* condition. This filtering narrows down the set of eligible bios to 38. The authors then independently screened the resulting 38 bios including the corresponding explanations, and assigned a rating of green ("in favor of using it"), yellow ("indifferent"), or red ("in favor of discarding it"), based on both a bio's content as well as the associated explanation, favoring bios that were non-trivial but that contained enough information to possibly make a correct prediction. We then decided on the final set of 14 bios based on majority vote, taking into account the required composition of scenarios, as outlined in Table 8.1 on page 180.

8.3.5 Measuring Reliance and Fairness

Measuring Reliance Behavior In our assessment of reliance behavior, we distinguish four cases, as depicted in Table 8.2 on page 182. We refer to cases where humans adhere to correct AI recommendations as *correct adherence*, to cases where humans adhere to wrong recommendations as *detrimental adherence*, to cases where humans override correct recommendations as *detrimental overriding*, and to cases where humans override wrong recommendations as *corrective overriding*. Note that the sum of shares of correct adherence and corrective overriding make up the final decision-making accuracy (Schöffer, Jakubik, et al., 2023). This taxonomy is similar to the one proposed by H. Liu et al. (2021) for trust; however, we want to stress the difference between trust and reliance. When comparing participants' reliance behavior across conditions, we compute and report the relative shares of any of these four types of reliance behavior on the 14 bios that participants see.

Measuring Distributive Fairness To evaluate distributive fairness of decisions, we measure disparities in error rates across gender (Barocas et al., 2019; Chouldechova,

Tab. 8.2.: Different types of reliance on AI recommendations.

	Human adherence to AI	**Human overriding of AI**
AI correct	Correct adherence	Detrimental overriding
AI wrong	Detrimental adherence	Corrective overriding

2017), which is closely linked to the ideas of *equalized odds* and *equal opportunity* (Hardt et al., 2016). From a fairness perspective, the goal is to minimize such disparities. We formalize them as follows: let FP_W be the share of wrongly predicted women professors, that is, women professors that are predicted to be teachers; and FT_W the share of wrongly predicted women teachers. Similarly define FP_M and FT_M for men. We can then quantify disparities in error rates as follows:

$$\text{Error rate disparity (Teacher} \rightarrow \text{Professor)} = |FT_W - FT_M|$$
$$\text{Error rate disparity (Professor} \rightarrow \text{Teacher)} = |FP_W - FP_M|,$$

where we use the notation of "Teacher $\rightarrow$ Professor" to indicate teachers that are wrongly predicted as professors, and vice versa for "Professor $\rightarrow$ Teacher." If we assume that the occupation of *professor* is associated with a higher societal status than *teacher*, we may also refer to cases of "Teacher $\rightarrow$ Professor" as *promotions*, and to "Professor $\rightarrow$ Teacher" as *demotions*. This will be important in the discussion of our findings.

Measuring Fairness Perceptions To measure fairness perceptions, we provide a brief introduction and then ask participants' agreement with three statements, measured on 5-point Likert scales from 1 ("fully disagree") to 5 ("fully agree"). We operationalize this in our questionnaires similar to Colquitt and Rodell (2015) as follows:

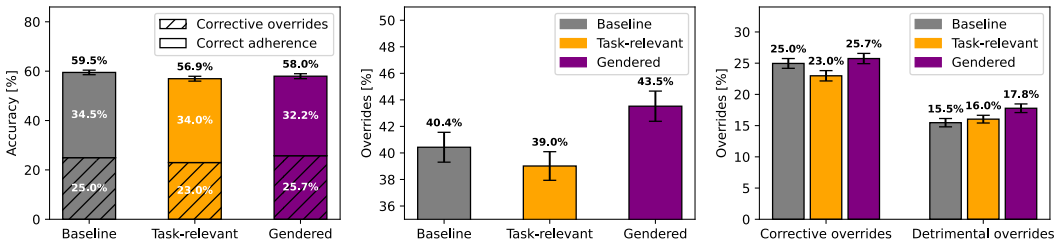
> *The questions below refer to the procedures the AI uses to predict a person's occupation. Please rate your agreement with the following statements.*
>
> (i) *The AI's procedures are free of bias.*
>
> (ii) *The AI's procedures uphold ethical and moral standards.*
>
> (iii) *It is fair that the AI considers the highlighted words for predicting a person's occupation.*

Note that items (i) and (ii) are taken from the *procedural justice* construct of Colquitt and Rodell (2015) and slightly rephrased to fit our case of AI-informed decision-making. These items have been frequently used in other human-AI studies, for instance, by Binns et al. (2018), Marcinkowski et al. (2020), and Schlicker et al. (2021). Colquitt and Rodell (2015) propose up to eight measurement items for procedural justice in the organizational psychology context; however, several of these items are not applicable here. Instead, we amend our questionnaires by a third item (iii) that is more tailored to our experimental setup. Since item (iii) is more explicit and we want to avoid priming, we ask (iii) last and without possibility to modify responses for (i) and (ii) retroactively. To obtain a single measure of fairness perceptions per participant, we eventually average ratings across the three items per participant; and we also confirm scale reliability (see Section 8.4.3).

8.3.6 Data Collection

Our study has received clearance from an institutional ethics committee. Participants were recruited via Prolific (Palan & Schitter, 2018). We required participants to be at least 18 years of age, and to be fluent in English. We also sampled approximately equal amounts of men and women; no other pre-screeners were applied. After consenting to the terms of our study, participants were then randomly and in equal proportions assigned to one of our three conditions and asked to complete the respective questionnaire. Overall, we recruited 600 lay people through Prolific. At the time of taking the survey, 13.5% of participants were 18–24 years old, 32.6% were 25–34 years old, 21.3% between 35–44, 13.8% between 45–54, 11.3% between 55–64, and 7.6% were older than 65. Regarding gender, 49.2% identified as women, 48.0% as men, and 1.8% identified as non-binary, third gender, or preferred not to say. 8.0% of participants are of Spanish, Hispanic, or Latinx ethnicity; and the majority (78.4%) considered their race to be White or Caucasian, followed by Black or African American (7.0%) and Asian (6.1%). For their participation, participants were paid on average £10.58 (approximately $12.70 at the time the study was conducted) per hour, excluding individual bonus payments of £0.05 per correctly predicted occupation. Participants took on average 10:12min (baseline), 12:51min (*task-relevant*), and 12:27min (*gendered*) to complete the survey.

Fig. 8.3.: Comparison of accuracy, total overrides, and types of overrides across conditions.



(a) Accuracy (b) Total overrides (c) Types of overrides

Note: We provide standard errors as error bars, where we compute the measure of interest (e.g., accuracy) for each individual participant in a given condition, then compute the standard deviation across all participants in that condition, and divide the result by the square root of the number of participants in that condition.

8.4 Analysis and Results

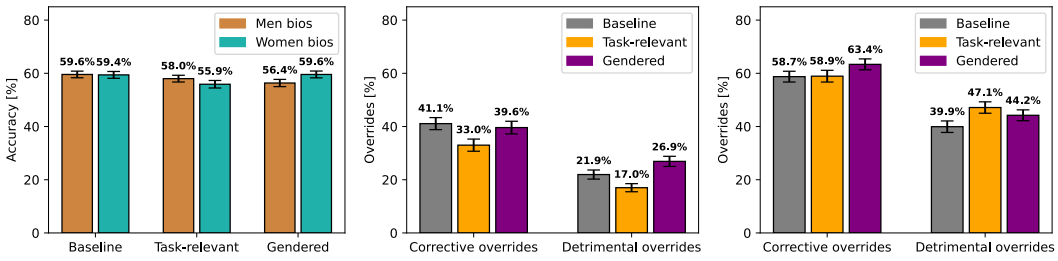
We first present results on the effects of explanations on accuracy as well as overriding behavior. Then, we examine how reliance behavior translates to distributive fairness. Finally, we assess the role of fairness perceptions. For all statistical comparisons, we conduct nonparametric tests because we cannot confirm the prerequisites (normal distribution and equal variance) of their parametric counterparts. Specifically, we conduct Kruskal-Wallis omnibus tests (Kruskal & Wallis, 1952) whenever applicable, and two-tailed Mann-Whitney U tests (Mann & Whitney, 1947) for pairwise comparisons.

8.4.1 Effects of Explanations on Accuracy and Overriding Behavior

Effects on Accuracy First, we examine how accuracy may be different between the baseline and the conditions with explanations, *task-relevant* and *gendered*. Mean accuracies[3] per condition are $M_{base} = 59.49\%$ ($SD_{base} = 13.11$), $M_{rel} = 56.94\%$ ($SD_{rel} = 13.86$), and $M_{gen} = 57.96$ ($SD_{gen} = 14.30$), as shown in Figure 8.3 (a). The Kruskal-Wallis omnibus test further suggests that there are no significant differences between the three means ($p = 0.260$). Recall that participants were incentivized through bonus payments to accurately predict occupations. This suggests that explanations did not aid human-in-the-loop decision-making when measured in terms of accuracy.

[3]We use M as a shorthand for *mean*, and SD for *standard deviation*. We also use the subscripts $base$, rel, and gen to refer to the baseline, *task-relevant*, and *gendered* conditions, respectively.

Fig. 8.4.: Comparison of accuracy by gender, and overriding behavior for men and women bios across conditions.



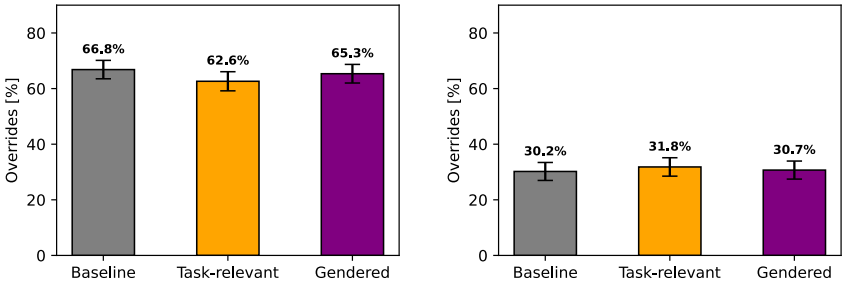
(a) Accuracy by gender (b) Overriding (men bios) (c) Overriding (women bios)

Effects on Overriding Behavior In Figure 8.3 (b) and (c) on page 184, we see that participants in the *gendered* condition overrode more AI recommendations than in the *task-relevant* condition ($p = 0.005$), and marginally more than the baseline ($p = 0.082$). From Figure 8.3 (c) on page 184, we further conclude that *both* corrective *and* detrimental overrides are highest in the *gendered* condition, with detrimental overrides being significantly higher than the baseline ($p = 0.012$). We interpret this increase in overrides further in Section 8.4.2. In the *task-relevant* condition, we see that overall overrides are lowest across conditions, with corrective overrides marginally decreasing ($p = 0.097$) over the baseline (see Figure 8.3 (c) on page 184). Overall, we conclude that people's reliance behavior is affected by how the AI system explains its recommendations; notably, people overrode AI recommendations more often when explanations highlight features that are evidently associated with gender. Across conditions, we also infer from Figure 8.3 (c) on page 184 that participants generally performed more corrective than detrimental overrides, and that the ability to perform corrective versus detrimental overrides did not improve through the provision of explanations.

8.4.2 Interplay Between Explanations, Reliance, and Distributive Fairness

Accuracy by Gender Consistent with our findings at the aggregated level (see Figure 8.3 (a) on page 184), we do not observe any accuracy changes through explanations over the baseline in Figure 8.4 (a), neither for men ($p = 0.199$) nor women ($p = 0.151$) bios. This means that both in the *task-relevant* and the *gendered* condition, explanations did not enable people to improve decision-making accuracy, neither for men nor women bios.

Fig. 8.5.: Detrimental overrides of AI recommendations that predict men teachers to be teachers (MTT) and women professors to be professors (WPP).



(a) Overrides for MTT

(b) Overrides for WPP

Types of Overrides by Gender and Occupation When looking at effects of explanations on overriding behavior by gender in Figure 8.4 (b) and (c) on page 185, no intervention improved participants' ability to perform corrective versus detrimental overrides of AI recommendations compared to the baseline, neither for men nor women bios. This is consistent with our findings at the aggregate level (see Figure 8.3 (c) on page 184). Notably, we see that in the *gendered* (Figure 8.4 (b) on page 185) and the *task-relevant* (Figure 8.4 (c) on page 185) condition detrimental overrides marginally increase over the baseline ($p = 0.078$ and $p = 0.013$), whereas corrective overrides remain unchanged.

From Figure 8.4 (b) and (c) on page 185 we also see that participants generally overrode more recommendations for women than men bios. However, this is not due to gender: we show in Figure 8.5 that across conditions there are more overrides for men teachers predicted by the AI system as teachers than for women professors predicted as professors. Together, these results suggest that people were overall more prone to do promoting[4] overrides; which means that participants overrode AI recommendations more often when someone was suggested to be a teacher versus a professor.

Importantly, people's likelihood to override conditioned on gender and predicted occupation did vary across conditions. By virtue of our study design, we are able to observe stereotype-countering[5] corrective overrides, and both stereotype-aligned and stereotype-countering detrimental overrides. As explained in Section 8.3.2, the motivation for this design is our focus on studying whether explanations allow

[4]We assume here that the occupation of *professor* is associated with a higher societal status than that of *teacher*. Hence, *promoting* refers to predicting a teacher to be a professor, whereas *demoting* means to predict a professor to be a teacher.

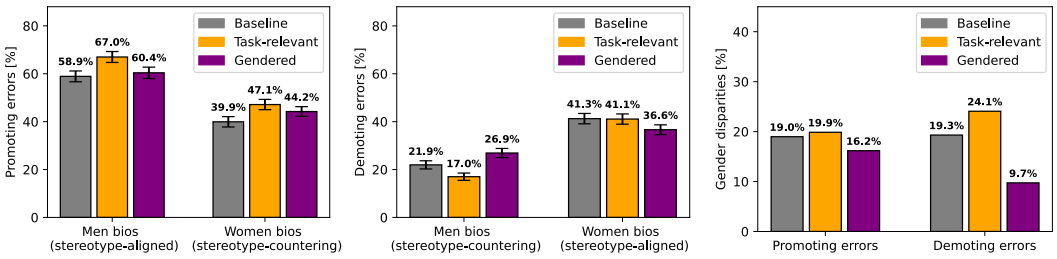
[5]Recall that societal stereotypes typically associate men with being professors and women with being teachers (J. Miller & Chamberlin, 2000).

humans to correct for stereotype-aligned wrong AI predictions, which would be the most frequent errors of an occupation prediction model that exhibits gender bias (De-Arteaga et al., 2019). We see that in the *task-relevant* condition, people perform fewer corrective overrides for men ($p = 0.011$) and the same amount for women ($p = 0.834$) in comparison to the baseline, as shown in Figure 8.4 (b) and (c) on page 185. Meanwhile, in the *gendered* condition participants perform marginally more corrective overrides for women ($p = 0.083$) and the same amount of such overrides for men ($p = 0.588$). This means that participants in the *gendered* condition were more likely to perform stereotype-countering corrective overrides than in the baseline, while participants in the *task-relevant* condition were less likely to do so.

As for detrimental overrides, we see that they marginally increase in the *gendered* condition for both men ($p = 0.078$) and women ($p = 0.110$) bios, compared to the baseline (Figure 8.4 (b) and (c) on page 185). Considering that we do not observe differences in stereotype-aligned detrimental overrides between conditions (see Figure 8.5 on page 186), we infer that people in the *gendered* condition performed more stereotype-countering detrimental overrides, by predicting more men to be teachers and women to be professors. It is noteworthy that when contrasting corrective and detrimental overrides, we observe that no condition improved participants' ability to make stereotype-countering *corrective* overrides versus stereotype-countering *detrimental* overrides. In the *gendered* condition, this means that participants became more likely to override an AI recommendation when it predicted that a woman is a teacher, irrespective of her true occupation. Overall, we observe reliance behavior in the *gendered* condition that counters societal stereotypes, whereas in the *task-relevant* condition people tend to rely on AI recommendations in a way that reinforces stereotypes. We elaborate on the implications of this for distributive fairness below.

Implications for Distributive Fairness We now examine how the observed reliance behavior relates to distributive fairness with respect to disparities in errors between men and women. First, we note that in the baseline condition, people tend to make more errors that promote men versus women (58.9% versus 39.9% in Figure 8.6 (a) on page 188), and erroneously demote women more than men (41.3% versus 21.9% in Figure 8.6 (b) on page 188). Note that in the case of men, promoting behavior is stereotype-aligned, whereas in the case of women such behavior is stereotype-countering; and vice versa for demoting behavior. The resulting absolute error rate disparities between men and women for the baseline are, hence, 19.0% (promotions) and 19.3% (demotions), as depicted in Figure 8.6 (c) on page 188. From the previous paragraph we know that people in the *task-relevant* condition

Fig. 8.6.: Analysis of promoting and demoting errors, as well as disparities in such errors between genders.



(a) Promoting errors (b) Demoting errors (c) Gender disparities

Note: We calculate errors that the human-in-the-loop makes as a consequence of their different reliance on AI recommendations. Promoting errors in (a) are errors where a teacher is eventually predicted to be a professor, and demoting errors in (b) are the ones where a professor is eventually predicted to be a teacher. We disaggregate these analyses by gender and indicate whether they align with or counter societal stereotypes. In (c), we report the absolute differences between men and women bios, both for promoting and demoting errors.

showed a tendency of reinforcing stereotypes, meaning that promotions of men increased more than those of women, which increased disparities in promotions even further over the baseline (Figure 8.6 (c), left). Similarly, demotions of men decreased much more than demotions of women, leading to increased disparities in demotions over the baseline (Figure 8.6 (c), right). In conclusion, we note that people's stereotype-aligned reliance behavior in the *task-relevant* condition exacerbated existing disparities in the baseline condition and, hence, hindered distributive fairness.

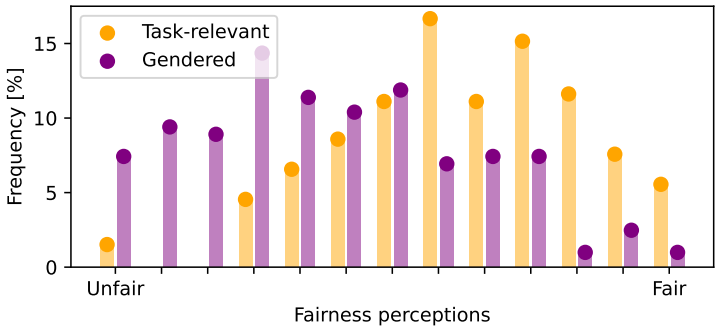
In the *gendered* condition, on the other hand, people countered stereotypes, meaning that promotions of women increased more than for men, reducing existing disparities (Figure 8.6 (c), left). The most drastic reduction in disparities happens for demotions (Figure 8.6 (c), right), since demotions *increased* for men and *decreased* for women (Figure 8.6 (b)). This results in a reduction of disparities in demotions from 19.3% (baseline) to 9.7% (*gendered* condition). Hence, people's stereotype-countering reliance behavior in the *gendered* condition mitigated existing disparities and, hence, fostered distributive fairness. It is important to stress that while disparities in error types decreased in the *gendered* condition compared to the baseline, this was due to a shift in the types of errors, as opposed to an increased ability to override mistaken AI recommendations. We report all p-values for pairwise comparison tests in Table 8.3 on page 189.

Tab. 8.3.: Results of pairwise comparisons.

Comparison	8.3 (b)	8.3 (c)		8.4 (b)		8.4 (c)		8.6 (a)		8.6 (b)	
		Corr.	Detr.	Corr.	Detr.	Corr.	Detr.	Man	Wom.	Man	Wom.
base – rel	0.307	0.097	0.374	0.011	0.047	0.834	0.013	0.011	0.013	0.047	0.834
base – gen	0.082	0.485	0.012	0.588	0.078	0.083	0.110	0.588	0.110	0.078	0.083

Note: We report p-values of two-tailed nonparametric Mann-Whitney U tests for pairwise comparisons. We provide p-values for both the comparison between baseline and task-relevant as well as baseline and gendered condition. Column names refer to the corresponding figures.

Fig. 8.7.: Distribution of fairness perceptions by condition.

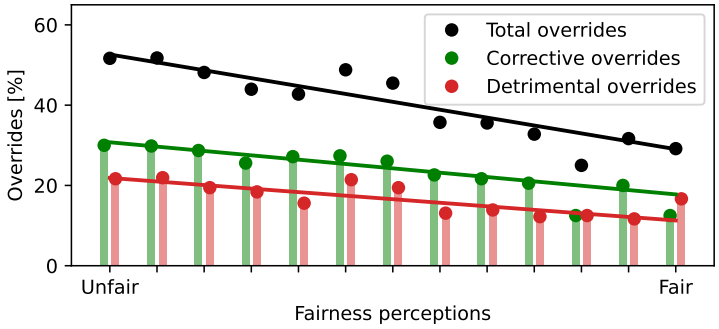


Note: Fairness perceptions are averages of three items measured on 5-point Likert scales, resulting in values between 1 (unfair) and 5 (fair) with 0.33 increments.

8.4.3 The Role of Fairness Perceptions

Effects of Explanations on Fairness Perceptions Recall that we measure three items regarding fairness perceptions on 5-point Likert scales, ranging from 1 (unfair) to 5 (fair), as outlined in Section 8.3.2. We then take the average of the three item ratings for each participant to obtain a single measure of fairness perceptions. We confirm good scale reliability at a Cronbach's alpha value of 0.77 (Taber, 2018). From Figure 8.7, we see that participants in the *task-relevant* and *gendered* conditions have significantly different perceptions of fairness towards the AI system. Concretely, we observe $M_{rel} = 3.53$ $(SD_{rel} = 0.85)$ in the *task-relevant* condition, and $M_{gen} = 2.54$ $(SD_{gen} = 0.98)$ in the *gendered* condition. This means that people who are shown a highlighting of task-relevant words perceived the underlying AI system as fairer than people who were shown gendered words as being important for given AI recommendations. Overall, we confirm prior works' findings and conclude that the AI system was perceived as significantly less fair when explanations point at the use of sensitive features compared to cases where explanations point at task-relevant features.

Fig. 8.8.: Relationship between fairness perceptions and overriding of AI recommendations.



Relationship of Fairness Perceptions With Overriding Behavior When we look at people's overriding behavior as a function of their fairness perceptions, we find an overall strong negative relationship ($p = 1.10 \cdot 10^{-11}$) between fairness perceptions and overriding of AI recommendations, that is, participants overrode the AI system more often when their fairness perceptions were lower. Concretely, we see that people overrode on average 52% of AI recommendations when their fairness perceptions were lowest, and only 31% when their fairness perceptions were highest. This negative relationship is consistent in both the *task-relevant* and the *gendered* condition, and it also persists when we disentangle corrective and detrimental overrides at the aggregate level. Figure 8.8 shows the relationship of overrides—both corrective, detrimental, and total—as a function of fairness perceptions for the *gendered* condition. Dots represent mean values of overrides for a given level of perceptions, and lines are OLS regressions fitted on the original data. All slopes in Figure 8.8 are significantly negative (total: $p = 1.97 \cdot 10^{-7}$; corrective: $p = 9.18 \cdot 10^{-5}$; detrimental: $p = 1.53 \cdot 10^{-4}$). We observe that as participants overrode more AI recommendations in the *gendered* condition, the rates at which corrective and detrimental overrides increase are approximately equal—in other words, the ratio of corrective to detrimental overrides is constant across perceptions. Overall, we conclude that people's fairness perceptions are associated with their reliance behavior in a way that low perceptions relate to more overrides than high perceptions. However, both corrective *and* detrimental overrides increased as fairness perceptions decreased. This implies that perceptions are not an indicator of people's ability to perform corrective versus detrimental overrides, but tend to only be associated with the quantity of overrides.

8.5 Discussion and Conclusion

Summary of Findings In this work, we conducted a first holistic analysis of the effects of feature-based explanations on distributive fairness in AI-informed decision-making. We also studied the mediating roles of reliance behavior and fairness perceptions, which have been the focus of prior work. Our findings suggest that feature-based explanations can have different effects on people's perceptions, their reliance behavior, and distributive fairness—depending on whether they highlight the use of task-relevant words or words that are proxies for sensitive attributes. Specifically, we observe that for the task of occupation classification, a highlighting of gendered words led to lower fairness perceptions, which are associated with more overrides of AI recommendations. On the other hand, when task-relevant words are highlighted, this led to higher fairness perceptions, which translate to fewer overrides. In no case, however, do we observe that explanations improve people's ability to perform corrective versus detrimental overrides, compared to a scenario with no explanations. Finally, we show that feature-based explanations can improve *or* hinder distributive fairness by fostering shifts in errors that counter or reinforce stereotypes: in the *gendered* condition, participants displayed stereotype-countering reliance behavior, while in the *task-relevant* condition, they displayed stereotype-aligned behavior. In both these cases, the respective reliance behavior affected both corrective and detrimental overrides. This means that the conditions affected the likelihood to perform an override conditioned on the predicted occupation and a bio's associated gender, but with no relationship to the true occupation. For instance, the *gendered* condition fostered *more* overrides of AI recommendations when a woman was predicted to be a teacher, irrespective of whether this prediction was correct. Meanwhile, in the *task-relevant* condition participants were *less* likely to override recommendations where a man was predicted to be a professor, irrespective of his true occupation.

Limitations Our study setup assigned participants to either the *gendered* or the *task-relevant* condition; that is, participants saw either only explanations with highlighting of gendered words or task-relevant words. We made this choice because we wanted to measure perceptions of fairness, but eliciting perceptions at an instance level could lead people to anchor their decisions to their expressed perceptions (or vice versa), which would compromise external validity. Assigning people to different conditions enabled us to measure perceptions at the aggregate level. In practice, an AI system might sometimes highlight only sensitive features, sometimes only task-relevant features, and at other times a mix of both. Future work that studies

how instance-level perceptions relate to aggregate-level perceptions, and how these interdependencies shape reliance behavior could complement our findings. While our study design does not explicitly account for this, even if perceptions vary at the instance level, our findings suggest that reliance would depend on the inclusion of sensitive features, which research has shown to be an unreliable signal for assessing fairness (Apfelbaum et al., 2010; Dwork et al., 2012; Kleinberg et al., 2018; Lakkaraju & Bastani, 2020; Nyarko et al., 2021; Pedreshi et al., 2008; Pruthi et al., 2020). In particular, previous research has shown that "fairness through unawareness," that is, the exclusion of information that is evidently indicative of a person's demographics, is neither necessary nor sufficient for an algorithm to be procedurally fair (Lakkaraju & Bastani, 2020; Nyarko et al., 2021; Pruthi et al., 2020) or to not display bias in terms of distributive fairness (Apfelbaum et al., 2010; Dwork et al., 2012; Kleinberg et al., 2018; Pedreshi et al., 2008). Our work complements these works by showing that feature-based explanations may foster stereotype-aligned reliance behavior, therefore *hindering* distributive fairness of AI-informed decisions.

Importantly, our study does not claim that the observed effects will necessarily generalize beyond the given setup. Instead, with this work, we aim to provide an important example that shows how unreliable feature-based explanations are when it comes to effects on humans' reliance behavior and distributive fairness. Our hope is that this work will inform improved assessment and design of transparency mechanisms, leading to a nuanced understanding of when and how certain types of explanations can enable humans to improve fairness properties of a system.

Implications and Outlook A main argument of our work is that claims around explanations fostering distributive fairness must directly measure the impact of explanations on fairness metrics of AI-informed decisions, which depend on humans' reliance behavior. To this end, our study constitutes a blueprint that should be used to evaluate other types of explanations and tasks. Crucially, our research shows that the mechanism through which reliance behavior affects metrics of fairness matters. In particular, we show that distributive fairness may improve even in the absence of an enhanced ability to perform corrective overrides. In other words, the presence of explanations may drive a change in fairness metrics by fostering over- or under-reliance for certain types of cases. This finding may be particularly important from a design and a policy perspective, since a common motivation when providing humans with discretionary power to override decisions is an expectation that they will be able to correct for an AI system's mistakes (De-Arteaga et al., 2020; European Union, 2016).

These findings also have implications for the interpretation of studies focused on perceptions of fairness (Starke et al., 2022). Our work shows that fairness perceptions have no bearing on people's ability to correctively override AI recommendations. Instead, our study results suggest that low fairness perceptions are associated with more overrides of AI recommendations, irrespective of their correctness. This may still lead to improvements in distributive fairness but does not indicate that humans differentiate between correct and wrong AI recommendations. This is important as perceptions are often used as proxies for trust and reliance (Starke et al., 2022).

Previous work has emphasized that transparency is not a monolithic concept, and that the design of explanations should always be grounded on a concrete objective that it helps advance (Lipton, 2018). Our work emphasizes the importance of designing explanations with the explicit purpose of enabling people to rely on AI recommendations in a way that enhances distributive fairness, and it casts doubt over the reliability of popular transparency mechanisms to advance this goal. To this point, novel findings from ethnographic work studying the use of AI systems have the potential to inform alternative designs of explanations. For instance, Lebovitz et al. (2022) study the adoption of AI systems in the healthcare domain and emphasize the importance of *interrogation practices*, which are practices used by humans to relate their own knowledge to AI predictions. Other works have studied interventions that help humans reason over the information that is and is not available to the AI system (Hemmer et al., 2022; Holstein et al., 2023). Future studies should explore whether explanations of the broader sociotechnical system better enable humans to perform corrective overrides that foster distributive fairness.

Part IV

Conclusion

9 Summary of Findings

Issues concerning transparency and fairness in artificial intelligence (AI)-informed decision-making have become increasingly apparent. There is a prevailing assumption that these subjects are closely intertwined, yet the exact dynamics of their relationship remain elusive. This thesis seeks to illuminate the relationship between transparency and fairness through a series of empirical and theoretical contributions. In this chapter, we consolidate our findings and revisit the initial research questions that guided our work. We will reiterate each research question and discuss how the results of this thesis contribute to their resolution. We will delve into the implications of these findings in Chapter 10, where we also contextualize them within the broader academic discourse. Finally, in Chapter 11, we provide a forward-looking perspective and conclude this thesis.

Research Question RQ1
What are the desiderata for transparency mechanisms with respect to fairness in AI-informed decision-making?

First, in Chapter 3, we carried out a structured literature review on the interplay between transparency and fairness in AI-informed decision-making. We discovered that a significant portion of previous work views transparency as a facilitator for fairness. Specifically, our qualitative analysis led us to identify several key assertions commonly found in the literature. Some posit that transparency is a prerequisite for fairness, implying that without transparency, AI-informed decision-making cannot be fair. Another perspective argues that transparency is not just necessary, but also sufficient for fairness, suggesting that adequate transparency automatically ensures fairness. It has also been proposed that transparency mechanisms, such as explanations, enhance stakeholders' perceptions of fairness. Finally, several claims relate to transparency as a tool that enables humans to assess, analyze, mitigate, and certify fairness.

In general, a considerable portion of previous work has expressed optimism about the role of transparency as a catalyst for fairness. However, our analysis of these claims also revealed that many of them are based more on intuition than empirical

evidence, potentially amounting to wishful thinking. In fact, many of these desiderata have not been adequately studied—and when they have, the results are often inconclusive. Furthermore, we identified two arguments that depict transparency as a potential threat to fairness. Some studies suggest that transparency mechanisms are susceptible to misinterpretation and deception, and that transparency and fairness may be conflicting objectives, indicating a possible trade-off between them. These insights informed the remainder of this thesis, immediately leading to the next research question, asking whether and how transparency and fairness can be reconciled in practice.

Research Question RQ2
How can we design AI systems that are inherently transparent and fair?

In Chapter 4, we showed that transparency and fairness are not necessarily contradictory objectives. AI systems traditionally depend on labeled data to train a classifier. However, in many situations, we lack access to ground-truth labels and must rely on labels derived from human decisions, which may be biased. Despite potential biases, these historical decisions often provide certain valuable insights into the true, unobserved labels—for instance, on the relative importance of legitimate features. With these observations in mind, we proposed a novel, fairness-aware AI system that employs a ranking-based approach. This system is grounded in monotonic relationships between legitimate features and the outcome, a key factor in making AI systems inherently transparent (Molnar, 2020). More precisely, we introduced a distance-based decision criterion, which utilizes legitimate information from past decisions and accounts for unwanted correlations between protected and legitimate features. In doing so, features with a high correlation to sensitive information have less influence on the final decision.

Our comprehensive experiments, conducted on both synthetic and real-world data, demonstrate that our method upholds a notion of meritocratic fairness by (i) assigning the desirable outcome to the most qualified individuals, and (ii) eliminating the influence of stereotypes in decision-making. Consequently, our method outperforms traditional supervised machine learning algorithms on several pertinent fairness metrics. Furthermore, we provided theoretical evidence that our method aligns with a well-established concept of individual fairness, which posits that similar individuals should be treated similarly (Dwork et al., 2012). We recognize, however, that the challenges associated with transparency and fairness extend beyond the scope of

purely technical solutions. Consequently, we also set out to study human perceptions towards AI systems that are used to inform consequential decisions.

> **Research Question RQ3**
> How do transparency mechanisms affect people's fairness perceptions towards AI systems?

In Chapter 5, we carried out a mixed-method online experiment to evaluate individuals' perceptions of *informational fairness*—that is, the belief that they are provided with sufficient information about a given decision-making process and its outcomes—and *trustworthiness* of an AI system when presented with varying levels of information about the system. We implemented an AI system in the context of automated loan approval and created various explanations that are popular in both research and practice. In our between-subjects conditions, we then manipulated the amount of information that participants saw in conjunction with exemplary scenarios of rejected loans.

Our quantitative analysis suggests that the amount of information provided and individuals' self-assessed AI literacy significantly impact their perception of informational fairness: both more information and higher AI literacy are associated with higher fairness perceptions. The informational fairness perceptions, in turn, positively correlate with the perceived trustworthiness of the AI system. These findings have important implications for designing and assessing transparency mechanisms, which we elaborate on more thoroughly in Chapter 10. Through an in-depth analysis of qualitative feedback, we also elicited what people desire from explanations. These desiderata include (i) consistency (both in line with people's expectations and across different explanations), (ii) disclosure of monotonic relationships between features and outcomes, and (iii) actionable recommendations, meaning that explanations should provide information relevant to reversing unfavorable decisions (i.e., recourse). Based on these findings, we were then interested in understanding how perceptions may be different between cases where an AI system versus a human being makes the final decision.

> **Research Question RQ4**
> How do people's fairness perceptions differ towards a human versus an AI system as the final decision maker?

In a follow-up study (Chapter 6) to the previous one, we conducted another online experiment to examine people's perceptions of fairness and trustworthiness towards AI systems in comparison to a scenario where a human instead of an AI system makes a high-stakes decision. Importantly, we provided identical explanations regarding decisions in both cases. Interestingly, our findings reveal that people perceive AI systems as more informationally fair than human decision makers. For instance, some study participants thought that automated decision-making is "fair by design," and others argue that "AI systems state the criteria and follow [them], there is no room for subjectivity and the data is used to make an objective decision." Our analyses also indicate that humans' AI literacy influences their perceptions, suggesting that those with higher AI literacy tend to favor AI systems more strongly over human decision makers. In contrast, individuals with lower AI literacy do not exhibit significant differences in their perceptions. Considering that humans are not just the subjects of decisions but can also be decision makers in a human-in-the-loop decision-making process, it is crucial to understand how their interaction with AI systems, particularly when transparency mechanisms are in place, impacts the quality of decisions. This led to our next research question.

Research Question RQ5
What is the relationship between human reliance on AI-based decision recommendations and common measures of decision quality?

In human-in-the-loop decision-making, a central promise of providing humans with discretionary power is that they should be able to complement the AI system by correcting its mistakes. In practice, however, we often see that humans tend to over- or under-rely on AI recommendations, meaning that they either adhere to wrong or override correct recommendations. In Chapter 7, we mathematically articulated and analyzed this interdependence between reliance behavior and accuracy in AI-informed decision-making. To this end, we developed a taxonomy on reliance behavior along two axes: (i) humans can *adhere to* or *override* AI recommendations, and (ii) AI recommendations can be *correct* or *wrong*. For humans to effectively complement an AI system, they need to adhere to the AI system's recommendations if and only if they are correct, and override them otherwise. Accuracy in human-in-the-loop setups is then defined as the proportion of instances where the human either adhered to a correct AI recommendation or overrode an incorrect one. Crucially, the capacity for humans to complement an AI system depends on three factors: the baseline AI accuracy (i.e., the initial quality of the AI recommendations), humans' reliance *quantity* (i.e., the frequency with which humans adhere to AI

recommendations), and reliance *quality* (i.e., humans' ability to discern between correct and incorrect AI recommendations). Even if the human decision maker can identify all mistakes, that does not necessarily mean that they can complement the AI system if they also override correct recommendations (*under-reliance*). Conversely, if the human-in-the-loop adheres to all correct recommendations, it can negatively impact accuracy if they also adhere to incorrect ones (*over-reliance*).

We also proposed a visual framework to make this interplay between reliance and accuracy more tangible. This framework is intended to be used for interpreting and comparing empirical findings, as well as to obtain a nuanced understanding of the effects of interventions (e.g., explanations) in AI-informed decision-making. Finally, we inferred several interesting properties from the framework: (i) when humans under-rely on AI recommendations, there may be no possibility for them to complement the AI system in terms of decision-making accuracy; (ii) when humans cannot discern correct and wrong AI recommendations, no such improvement can be expected either; (iii) interventions may lead to an increase in decision-making accuracy that is solely driven by an increase in humans' adherence to AI recommendations, without any ability to discern correct and wrong. Our proposed framework also served as a blueprint for the study addressing our final research question.

Research Question RQ6
How do transparency mechanisms affect distributive fairness in human-in-the-loop decision-making?

In Chapter 8, we conducted a third online experiment centered around occupation classification from short biographies. Herein, we investigated the impact of feature-based explanations on distributive fairness in human-in-the-loop decision-making. This analysis also allowed us to explore how any effects are mediated by humans' perceptions and their reliance on AI recommendations, thereby connecting to our previous findings on fairness perceptions and reliance behavior. Our results suggest that feature-based explanations can influence people's perceptions, their reliance behavior, and distributive fairness in different ways, depending on whether they emphasize the use of task-relevant features or features that serve as proxies for sensitive attributes. Specifically, for the task of occupation classification, we found that highlighting features related to gender (e.g., gender pronouns) led to lower fairness perceptions, which were associated with more overrides of AI recommendations. Conversely, when task-relevant features were highlighted, this led to higher

fairness perceptions, which resulted in fewer overrides. However, we did not observe that explanations improved people's ability to perform corrective versus detrimental overrides, compared to a baseline scenario without explanations.

Finally, we demonstrated that feature-based explanations can either enhance *or* impede distributive fairness by encouraging shifts in errors that either counter or reinforce societal stereotypes. Concretely, when explanations highlighted sensitive features, humans displayed stereotype-*countering* reliance behavior, whereas when only task-relevant features were highlighted, they displayed stereotype-*aligned* behavior. In both cases, the respective reliance behavior was independent of the correctness of AI recommendations. This ultimately implies that explanations influenced the likelihood of performing an override based on the predicted occupation and a biography's associated gender, but with no correlation to the true occupation. For instance, explanations highlighting sensitive features encouraged more overrides of AI recommendations when a woman was suggested to be a teacher, regardless of whether this prediction was correct. Meanwhile, when only task-relevant features were highlighted, participants were less likely to override recommendations where a man was predicted to be a professor, irrespective of his actual occupation. In conclusion, we showed that effects on distributive fairness are brittle, and that the ability of popular feature-based explanations to advance this goal relies on a human-in-the-loop operationalization of the flawed notion of "fairness through unawareness" (see Section 2.5.4). We discuss implications of our findings in more detail in the next chapter.

10 Implications

The importance of addressing issues surrounding transparency and fairness in artificial intelligence (AI)-informed decision-making is paramount and will become even more critical with the forthcoming implementation of the AI Act in the EU and similar regulations. In fact, non-adherence to such regulations is not only morally questionable but could also result in significant financial penalties (Madiega, 2021). And these challenges are far from resolved. For instance, a recent study from Stanford University revealed that most foundation models, which include cutting-edge AI systems like GPT-4 (Bubeck et al., 2023), would not meet the compliance standards set by the current draft of the EU AI Act (Bommasani et al., 2023). Therefore, both transparency and fairness must be integral considerations during the development and deployment of AI systems, particularly when they are used to inform high-stakes decision-making in sectors such as education, employment, public service, and law enforcement. These (and more) are all areas that the AI Act identifies as high-risk applications (Madiega, 2021). This underscores the relevance and urgency of our research. In this context, our findings carry multiple implications for researchers, practitioners, and policymakers. We categorize these implications by theme in the following sections.

10.1 On Fairness Desiderata of Transparency

In Chapter 3 of this thesis, we saw that transparency mechanisms are frequently heralded as a silver bullet in AI-informed decision-making, particularly in relation to fairness. However, we also found that the supporting evidence for this claim is tenuous, in part due to our incomplete understanding of the capabilities and limitations of existing transparency mechanisms. This implies that claims portraying transparency as an ethical panacea are misleading. This thesis underscores the complex and multidimensional relationship between transparency and fairness, suggesting that transparency may serve multiple functions in meeting fairness objectives. These different roles should be acknowledged in order to make informed claims about what we can and cannot reasonably expect from transparency mechanisms.

The first role that transparency serves is related to an *epistemic* aspect of fairness (Langer, Oster, et al., 2021). This aspect is satisfied if a stakeholder is empowered to evaluate whether an AI system is fair, which hinges on an individual's personal definition of fairness. For instance, if they believe that the use of sensitive information by an AI system is either fair or unfair, then a feature-based explanation can provide this insight by highlighting which features are considered in the decision-making process. Other notions of fairness may not be as easy—or even impossible—to assess. When evaluating the epistemic aspect, it is essential that the stakeholder's fairness assessment is calibrated (Schöffer & Kühl, 2021). This aligns with what Lazar (2022, p. 1) refers to as a "justified understanding." If stakeholders merely *believe* they can evaluate fairness without a grounded basis, this does not provide a meaningful criterion for assessing the effectiveness of a given transparency mechanism. We discuss this further in Sections 10.3 and 10.5.

The second aspect is a *substantial* one, which suggests that stakeholders have specific fairness properties they wish to see embodied in an AI system (Langer, Oster, et al., 2021). For instance, if a stakeholder desires the AI system to uphold demographic parity, an explanation would need to empower that stakeholder to ensure that the AI system indeed satisfies this particular notion of fairness. This is something that transparency mechanisms cannot directly provide. However, depending on a stakeholder's agency in the process, an explanation might *indirectly* enable them to intervene and shape the system's outcomes in ways that align with a specific notion of fairness. For instance, if a human-in-the-loop is presented with an explanation, it could empower them to take action and influence the fairness attributes of the AI system, based on the information conveyed by the explanation. However, our findings from Chapter 8 raise doubts about the reliability of existing transparency mechanisms to achieve this. We note that the epistemic and substantial facets of fairness desiderata can sometimes be correlated (Langer, Oster, et al., 2021). For instance, in Chapter 8, we found that fairness perceptions influenced how humans relied on the AI system, leading to different outcomes for the statistical fairness notion at hand. However, it is important to stress that the epistemic and substantial aspects are not inherently interconnected and need to be carefully disentangled. Simply put, being aware of unfairness does not automatically guarantee the achievement of fairness.

Lastly, explanations may possess an intrinsic fairness value, which is *deontological*, that is, agnostic of outcomes. This aligns with claims arguing that transparency is a sufficient condition for fairness, for instance, "the explanation of the decision process is a way to guarantee fairness to all people impacted by AI-related decision" (Ferreira & Monteiro, 2020, p. 2). This is also related to the concept of informational fairness,

which considers an AI system to be informationally fair if it provides adequate information about its processes and outcomes. We summarize the three facets of fairness desiderata, as they are of paramount importance in delineating the interplay of transparency and fairness:

Transparency may address three different facets of fairness desiderata in AI-informed decision-making:

- *Epistemic facet*: Ability of humans to assess fairness
- *Substantial facet*: Ability of humans to guarantee fairness
- *Deontological facet*: Transparency is sufficient for fairness

10.2 On Trade-Offs Between Transparency, Fairness, and Utility

It is sometimes argued that transparency and fairness are conflicting goals. In Chapter 4, we demonstrated that it is feasible to design AI systems that uphold both inherent model transparency and statistical notions of fairness. This shows that it is possible to implement effective guardrails to counteract the notion that simpler models are likely to result in unfair outcomes (Kleinberg & Mullainathan, 2019). However, our empirical results also indicate that the enforcement of transparency and fairness may compromise accuracy. From an optimization perspective, this is not surprising because introducing additional constraints reduces the size of the solution space (Bertsimas et al., 2011). Worryingly, decreases in accuracy can have financial implications for companies when accuracy is directly tied to the utility (e.g., profits) that an AI system generates (Schöffer, Ritchie, et al., 2023). This phenomenon, particularly in relation to fairness, has also been termed the "cost of fairness" (von Zahn et al., 2022).

However, the validity of accuracy as a measure of utility should be questioned when training data is heavily biased. In fact, the decrease in accuracy observed in our experiments in Chapter 4 is relative to test labels that are unjustifiably and significantly biased in favor of men. Consider an example from De-Arteaga et al. (2022), where a bank consistently underestimates women's financial standing and their likelihood of repaying a loan. This misestimation is detrimental not only to women but also to the bank itself, which misses out on potential profits associated

with these loans. This implies that in some contexts, enhancing fairness could lead to an *improvement* in utility, even if it results in a decrease in accuracy based on biased testing labels. In such scenarios, it might be beneficial to avoid using supervised machine learning or other inductive reasoning approaches that maximize for predictive accuracy altogether (Barocas et al., 2019). Instead, it could be more fruitful to employ AI systems similar to the one we propose in Chapter 4 of this thesis.

We also suggest that it is beneficial to evaluate any trade-offs between transparency, fairness, and accuracy differently for fully automated decision-making compared to human-in-the-loop setups. In the latter case, transparency mechanisms are typically aimed at human experts who strive to enhance the AI system's accuracy and fairness. In Chapter 8, we saw that the presence of post-hoc explanations may not enable the human-in-the-loop to improve accuracy beyond the AI system's baseline. We also observed that explanations can either increase or decrease distributive fairness. However, the nature of these trade-offs differs from those in fully automated systems. In fully automated systems, the trade-offs we typically refer to are related to the design and deployment of the AI system itself, and they are quantified on held-out testing data. In contrast, in human-in-the-loop decision-making, any trade-offs arise as a result of human intervention *after* an AI system has been deployed. Hence, reasoning about trade-offs requires novel frameworks that are tailored to human-in-the-loop setups. We address this shortly in Section 10.4.

10.3 On Measuring Human Fairness Perceptions

In Chapter 5, we observed that human perceptions of fairness are sensitive to variations in the amount of information provided, and that individuals have strong preferences concerning the appropriateness of certain features, such as gender or race, in decision-making. Additionally, in Chapter 6, we found that perceptions towards AI systems often rest on the assumption that these systems are inherently fair, which indicates that perceptions are frequently based on questionable assumptions. These findings align with previous research suggesting that human perceptions are brittle and can be easily manipulated through transparency mechanisms—in particular those that highlight the use or disuse of sensitive features, such as LIME or SHAP (Dimanov et al., 2020; Lakkaraju & Bastani, 2020; Slack, Hilgard, et al., 2020). Overall, we infer that transparency interventions can mislead individuals into forming uncalibrated perceptions, leading them to perceive AI systems as fair when they are not, or vice versa. This is particularly concerning as it could be

exploited by powerful stakeholders, such as system developers, to deceive more vulnerable stakeholders, like those that are affected by AI-informed decisions. Such deceptive practices, known as *dark patterns*, have long been a concern in user interface design (Gray et al., 2018) and are now also relevant to transparency mechanisms, as highlighted by Chromik et al. (2019). Our findings underscore this issue and emphasize the need to make transparency mechanisms resistant to exploitation.

This thesis also prompts us to consider when fairness perceptions are an end goal in themselves, and when it is crucial to understand the mediating role they play in relation to other downstream metrics. Given that perceptions can be easily misled and are often uncalibrated, it is vitally important to determine whether such uncalibrated perceptions also lead to uncalibrated *behavior*, such as under- or over-reliance on AI recommendations (Schöffer, De-Arteaga, & Kühl, 2022). There is limited research examining these downstream effects, and when such studies exist, they typically focus on the effects of fairness perceptions on other forms of perceptions, like trust or satisfaction (Starke et al., 2022) Our work in Chapter 8 begins to address this gap by demonstrating that fairness perceptions are related to the quantity, but not the quality, of reliance on AI recommendations. However, much more research is needed to holistically understand the role that fairness perceptions play in human-in-the-loop decision-making. To that end, the study design from Chapter 8 is intended to serve as a blueprint for similar future research in this area. Until we fully understand the relationship between perceptions and behavior, we must avoid conflating them, as has been done with trust and reliance (Lai et al., 2021). This leads us into our next theme of implications.

10.4 On Assessing Transparency Mechanisms

We, along with other researchers (e.g., Lipton (2018)), advocate for the necessity of grounding transparency mechanisms in the objectives they aim to achieve, and subsequently evaluating them against these goals. In the words of Lipton (2018, p. 23), we need "definitions of success" for transparency mechanisms; and these will depend on context. Specifically with respect to fairness, we observed in Chapters 3 and 8 that numerous assertions have been made regarding the role of explanations in promoting certain reliance behaviors that are beneficial to distributive fairness. However, most prior studies concerned with fairness have assessed the impact of explanations on fairness *perceptions* (Starke et al., 2022). This discrepancy between goal and evaluation may lead to incorrect conclusions and create unrealistic

expectations about the capabilities of transparency mechanisms. In human-in-the-loop setups, it is important to understand that the effects of explanations on fairness perceptions may not directly translate into the desired reliance behavior that is conducive to distributive fairness. Therefore, if one claims that explanations should promote specific notions of distributive fairness, then one needs to conceive and follow a research design that allows to measure precisely that.

When concerned with decision quality more broadly, we saw in Chapter 7 that transparency mechanisms in human-in-the-loop decision-making are typically evaluated with respect to their effects on (i) reliance behavior or (ii) decision-making accuracy. However, the definitions of *under-* and *over-reliance* are often ambiguous. Some researchers define these terms at the level of individual decisions, where *under-reliance* is considered as overriding a correct AI recommendation and *over-reliance* as adhering to an incorrect one (Schemmer et al., 2023). Others view under- and over-reliance as global behaviors, characterized by a general propensity of humans to override too many or too few AI recommendations, respectively. These terminological inconsistencies complicate the interpretation of statements related to explanations leading to over- or under-reliance. Our work in Chapter 7 provides guidance on how to interpret these terms, emphasizing that over- or under-reliance is not detrimental per se, depending on the baseline human reliance in the absence of an intervention. For instance, if humans tend to always accept AI recommendations in a control setting without explanations, then an intervention leading to more scrutiny and, hence, overrides may still be desirable from an accuracy standpoint—even if some of these overrides are of correct AI recommendations. Recognizing these nuances is crucial for meaningfully assessing the effects of transparency mechanisms, yet they have generally been overlooked in previous research.

Furthermore, our work shows that the implications of over- and under-reliance are not symmetrically related to accuracy. For instance, we illustrated that if an AI system's performance is high, accepting too few AI recommendations may never lead to human-AI complementarity with respect to accuracy. Conversely, when the human-in-the-loop is unable to distinguish between correct and incorrect AI recommendations, it is optimal in expectation to *always* accept AI recommendations, provided the system performs better than chance. This also implies that an observed increase in accuracy through a transparency mechanism could be solely due to people adhering to more AI recommendations, as opposed to an improved human ability to distinguish between correct and wrong AI recommendations. Relatedly, in Chapter 7, we demonstrated that there can be interventions that may appear identical when only assessing their effects on accuracy, but they might lead to drastically different reliance behavior. Overall, our work suggests that evaluating

transparency mechanisms solely based on accuracy can be misleading and may lead to uninformative or even deceptive conclusions. These are crucial considerations for empirical research assessing the effects of transparency mechanisms.

10.5 On the Effectiveness of Feature-Based Explanations

Finally, our research has several significant implications regarding the effectiveness of feature-based explanations, including popular methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), which are widely used in both academic and practical settings (Bhatt et al., 2020; ElShawi et al., 2021; Gilpin et al., 2018). First, as outlined in Section 10.3, we saw that these explanations can potentially mislead people's perceptions, suggesting that they may not be a reliable tool for helping individuals form accurate fairness perceptions of an AI system (Schöffer & Kühl, 2021). Second, as demonstrated in Chapter 8, we discovered that these explanations are not consistently reliable in terms of mitigating distributive unfairness in human-in-the-loop decision-making, despite prevailing beliefs to the contrary (Kite-Powell, 2022). Third, our research indicates that feature-based explanations do not generally enable humans to enhance the accuracy of AI systems, meaning that they do not enable them to differentiate between correct and wrong AI recommendations. This aligns with the conclusions drawn in several previous studies (Schemmer, Hemmer, Nitsche, et al., 2022).

Let us delve deeper into why these explanations have not been able to live up to their promises. In line with the three facets of fairness desiderata discussed in Section 10.1, it is crucial to differentiate between what we can and cannot reasonably expect from feature-based explanations. To this end, we may want to analyze the cues these explanations convey to their users (Schlicker et al., 2022). Generally, the concept of feature-based explanations revolves around identifying which features are deemed important in predicting a specific outcome (Ribeiro et al., 2016). However, this emphasis becomes fragile in the presence of redundant encodings (Dwork et al., 2012), meaning that sensitive features can be inferred from other, seemingly legitimate, features. Prior research has also demonstrated that through relatively simple interventions, it is possible to make a feature-based explanation highlight features that do not appear problematic, while the classifier's behavior on input data remains unchanged—and potentially unfair (Dimanov et al., 2020; Slack, Hilgard, et al., 2020). This casts doubt on the ability of feature-based explanations to reliably

enable humans to discern whether and how an AI system takes into account sensitive information.

Moreover, if feature-based explanations were to assist humans-in-the-loop in complementing an AI system in terms of (i) accuracy or (ii) distributive fairness, it would imply that these explanations must supply pertinent cues to do so. Specifically, they would need to enable humans to differentiate between (i) correct and wrong, and (ii) fair and unfair AI recommendations, and to utilize their discretionary power to advance these metrics. However, our research indicates a fundamental mismatch in relevance between the cues provided by these explanations and those required to reliably promote these desiderata. This has implications for system developers, suggesting that they should consider the cues an explanation provides, and whether they are relevant for achieving a specific goal (e.g., distributive fairness). Given these limitations of widely-used transparency mechanisms, we also propose that a simple "right to explanation" (Goodman & Flaxman, 2017), as is often advocated by policymakers, is not sufficient to ensure the responsible use of AI systems in decision-making.

11 Outlook

In this thesis, we have demonstrated that the relationship between transparency and fairness is complex and multifaceted. Transparency mechanisms are frequently touted as an ethical panacea in artificial intelligence (AI)-informed decision-making. However, through a series of theoretical and empirical contributions, we have shown that the reality is more nuanced, and that existing transparency mechanisms may not be as effective as anticipated. While our findings suggest that there are no universal solutions for either fairness or transparency, we do not wish to foster discouragement. Significant problems seldom have simple solutions. In this chapter, we expand our viewpoint and discuss a variety of potential avenues for progress.

Using Appropriate Baselines In terms of fairness, it is acknowledged that no technical solution can serve as a definitive guarantee of fairness. While it is easy to become paralyzed and conclude that fairness is an unsolvable problem, we wish to encourage and argue that this may not be the correct question to ask. The issue is not whether we can completely solve the problem of unfairness, but whether we can improve the situation compared to a relevant baseline. This perspective has also been recently advocated by De-Arteaga et al. (2022). In the context of decision-making, this baseline often involves a human decision maker, who in some instances has been shown to exhibit even more bias than AI systems; for instance, in the realm of clinical decision-making (Ganju et al., 2020). If previous practices were unfair, it may be unrealistic to expect AI systems to solve all problems. Furthermore, while the array of different fairness notions may seem overwhelming, in many cases it is not as challenging to justify why a particular notion of fairness should be upheld. In such instances, the Fairness, Accountability, and Transparency (FAccT) community has provided a wealth of tools to address issues surrounding unfairness. If we can convincingly argue that, for instance, the notion of predictive parity is worth upholding in risk assessment scenarios, then we have a variety of tools at our disposal to address this. Guidance on selecting appropriate fairness metrics for specific contexts exists, as illustrated by De-Arteaga et al. (2022).

Similar arguments can be made regarding transparency. Perhaps our expectations for transparency mechanisms are too high, especially considering that human expla-

nations are not flawless either. This suggests that we should perhaps recalibrate our expectations. In a similar vein, Zerilli et al. (2019, p. 1) question whether we hold AI system transparency to an "unrealistically high standard," essentially arguing that human decision-making is often not transparent either.

Grounding Interventions in Measurable Objectives When enforcing a fairness constraint in the development of an AI system, it is clear that our objective is to satisfy this particular notion of fairness. This clarity is not always given with transparency, where the goal can often be more nebulous and may even encompass an intrinsic value, independent of an outcome (Hayes, 2020). Such *deontological* accounts of transparency contrast with a *consequentialist* perspective, which assesses things "solely by the states of affairs they bring about" (Alexander & Moore, 2020, p. 1). However, even deontologists often assume that transparency at the very least fosters some type of "acquisition of knowledge" (Hayes, 2020, p. 1). Moreover, expecting no measurable benefit from transparency renders such mechanisms unassessable. Therefore, we should initially ask what we expect from an explanation, and then base the design of transparency mechanisms on quantifiable objectives. In this thesis, we have observed that transparency is hoped to fulfill a broad array of fairness-related desiderata, many of which are poorly defined. Langer, Oster, et al. (2021) assess the goals of transparency more generally and identify 29 objectives, such as privacy or responsibility, where it is not immediately evident how effectiveness could be measured in the first place. The necessity of anchoring transparency mechanisms in measurable objectives has been underscored in this thesis, and has also been highlighted in other research (Lipton, 2018). In this regard, the FAccT and Explainable AI (XAI) communities can greatly benefit from research in the social sciences and humanities, which have long engaged with the concept of explanations including their potential objectives more broadly (Lombrozo, 2012; T. Miller, 2019).

Our research also underscores the importance of designing appropriate experiments to empirically evaluate such objectives of transparency. For instance, if we desire a transparency mechanism to empower human decision makers to assess the fairness of an AI recommendation, then the explanation must convey the relevant cues. This has also been recently highlighted by Schlicker et al. (2022) in the context of trustworthiness. However, even if an explanation provides relevant cues, we must empirically verify whether humans can effectively utilize them. Based on the majority of empirical findings to date, there appears to be a discrepancy along this path, which must be considered and rectified in the design process of novel interventions.

Widening the Scope of Transparency It is recognized that transparency, much like fairness, is not a monolithic concept (Lipton, 2018). Just as we cannot resolve fairness issues with a single statistical metric, we cannot address transparency problems with a single style of explanation. Therefore, we believe it is not productive to consider transparency within rigid frameworks that dictate what does and does not constitute an explanation. What is needed is a shift from static, one-off explanations towards a broader understanding of what transparency might encompass. We should aim to understand how to build a supportive ecosystem around AI systems, one that enables each stakeholder to achieve their respective goals. One element of such an ecosystem could be an interface that allows individuals to query different pieces of information, based on their background and situational needs. In this context, insights from ethnographic work studying the use of AI systems have the potential to inform alternative explanation designs. Lebovitz et al. (2022), for instance, examine the adoption of AI in three healthcare domains and highlight the importance of interrogation practices, which are methods used by humans to connect their own knowledge to AI predictions. They note that if AI systems are to add value, they will sometimes make recommendations that conflict with experts' knowledge. Therefore, what is needed are processes and tools that assist them in reconciling these differing perspectives.

Moreover, it is not always clear that what is needed are explanations pertaining to the AI system's inner workings, as opposed to explanations of the broader sociotechnical system. For instance, interventions that assist humans in reasoning about the information that is and is not available to the AI system may help them reconcile disagreements and better integrate multiple information sources (Hemmer et al., 2022; Holstein et al., 2023). In clinical decision-making, Ehsan et al. (2023) discovered that transparency mechanisms could foster social interactions and reveal how different clinicians responded to specific AI recommendations in the past. Auxiliary interventions such as cognitive forcing functions have also been demonstrated to encourage more effective reliance behavior (Buçinca et al., 2021). Finally, in some situations, it may be best to provide no explanation at all, such as in circumstances where they could be used to—deliberately or inadvertently—deceive vulnerable stakeholders (Molnar, 2020).

Embracing Human-Centered Approaches Much of the work on fairness and transparency is technical in nature, focusing on the development of technical artifacts aimed at addressing issues of unfairness and opacity in AI-informed decision-making. While this work is crucial, it is not sufficient in the quest for responsible AI. Just as a single statistical fairness notion cannot certify an AI system's fairness, we need to

transcend the idea of a single explanation catering to the unique objective of one stakeholder. We argue that the task for the field of XAI is not merely to create the most compelling explanation that fosters maximum trust with the end user. Rather, the challenge lies in recognizing that a single explanation cannot meet the needs of all stakeholders, with all their diverse backgrounds and incentives. Therefore, we should focus on the question of *who* we are designing these AI systems for—*who* are the humans interacting with these systems? Ehsan et al. (2023, p. 1) refer to this as a "sociotechnical gap—[a] divide between the technical affordances and the social need" of explainable AI systems. To understand and address this gap, it is crucial to comprehend the goals and incentives of relevant stakeholders, which must be elicited through user studies, rather than making assumptions that are prone to misconceptions (T. Brown, 2008).

It is not only important to understand people's pain points and what they desire from transparency, but also to acknowledge differences in the range of actions available to all involved stakeholders. System developers have substantially more power in the process of designing and deploying AI systems than vulnerable groups like decision subjects. As such, we risk prioritizing the goals of powerful stakeholders over those of other stakeholders, which may often be in conflict with each other. For instance, system developers may be interested in fostering trust in their AI systems, whereas decision subjects are concerned with being treated fairly (Langer, Oster, et al., 2021). It is important to understand how such conflicts arise in a given context and what can be done to reconcile them. For any given AI system, we must critically ponder whose interests are being prioritized in the deployment of transparency mechanisms, and how this reflects power imbalances. If this is not done, we risk burdening explanations with conflicting desiderata, resulting in deficient and sometimes even harmful user experiences. To prevent the exploitation of conflicts of interest, we must establish processes of proper oversight and ensure the truthfulness of explanations. This should also include mechanisms like appropriate documentation (Gebru et al., 2021). In certain aspects, however, conflicts between stakeholders are too fundamental to be likely resolved, for instance, when AI systems involve intellectual property that must not be revealed. In such cases, Lazar (2022, p. 2) argues that transparency may be a lost cause because "one cannot reach a justified understanding of a secret."

Fostering Effective Human-AI Collaboration Most AI systems deployed for informing consequential decisions are not fully automated. Humans typically retain discretionary power in many critical domains, which underscores the need to better

understand how these humans perceive and interact with AI systems. In these scenarios, transparency is generally intended to offer more than just an understanding of the underlying AI system and its outcomes. Specifically, in human-in-the-loop decision-making, transparency mechanisms are often viewed as means of decision support for the human expert. As such, they are intended to provide the human with relevant cues to complement an AI system towards improved and fairer decisions. Here, the value of transparency transitions from mere information provision to guiding substantial interventions (van Berkel et al., 2019), indicating a shift from being descriptive to being *prescriptive*, in a way that grants the human control over the system. The human desire to obtain actionable recommendations through transparency is something that we have observed in Chapter 5 of this thesis as well. Therefore, we need to concentrate on designing interventions that reliably provide human experts (e.g., doctors, judges, or human resources professionals) with the information they need to complement AI systems by interfering when these systems make incorrect or unfair predictions. Crucially, the notion of fairness in human-in-the-loop decision-making is complex, encompassing the fairness of (i) the AI system, (ii) the human, and (iii) the interaction between the two. For instance, it would be essential to understand the circumstances and mechanisms through which biases of AI systems and humans are likely to offset or even amplify each other. While our research contributes several insights towards a better understanding of this intricate dynamic, there is a clear call for additional conceptual and empirical exploration in this area.

Finally, it is also crucial to explore different paradigms of operationalizing joint human-AI decision-making. A relevant line of research focuses on the idea of capitalizing on the individual strengths of AI systems and human experts (Hemmer et al., 2023; Madras et al., 2018; Mozannar & Sontag, 2020). The general premise hereby is that the entity best suited to make a given decision should be the one in control. To gauge the appropriate degree of human involvement, it is generally necessary to obtain a reliable estimate of an AI system's decision-making uncertainty. For instance, if an AI system's calibrated confidence in a prediction is very high, then more weight should be given to this recommendation than in cases where the system is relatively uncertain. If a system's prediction confidence falls below a critical threshold, a sensible response may be to refrain from issuing a prediction at all, and instead defer entirely to the human; especially when the human is an expert with contextual knowledge or other relevant capabilities. Such *learning-to-defer* (Madras et al., 2018) approaches have been deployed with much success, for instance, in breast cancer screening, where an AI system automatically assesses cases of high certainty and defers others to a radiologist (Leibig et al., 2022). It would

be interesting to adapt and test such setups with a focus on fairness—perhaps in conjunction with explanations that can indicate why an AI system ended up making a prediction or deferring to the human expert.

This thesis was conceived during a period of profound transformation in the AI research field. Novel AI systems like ChatGPT are integrating into our daily lives in unprecedented ways (K. Hu, 2023). While these systems present numerous opportunities, the associated ethical challenges, including biases and lack of transparency, must not be ignored (Bubeck et al., 2023). In fact, the latest *2023 State of AI* report from Stanford University notes a steady increase in AI controversies over recent years (Lynch, 2023). This suggests that issues related to transparency and fairness are far from resolved, and may indeed become even more critical in the future. This thesis aims to serve as a foundation for addressing some of these disconcerting trends, hoping that our findings can contribute to the responsible design and deployment of AI systems that can ultimately assist us in building a more equitable society.

Bibliography

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In *Human and machine learning* (pp. 21–35). Springer. (Cit. on pp. 57, 69, 266, 267).

ACTICO. (2021). Automated credit decisioning for enhanced efficiency [https://www.actico.com/blog-en/automated-credit-decisioning-for-enhanced-efficiency/ (Accessed: 2023-06-17)]. *ACTICO GmbH*. (Cit. on pp. 114, 136, 140).

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160 (cit. on pp. 7, 30, 32, 64, 72, 109, 115, 136, 137, 141, 170, 267).

Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology*, *154*(11), 1247–1248 (cit. on pp. 42, 43).

Adomavicius, G., & Yang, M. (2022). Integrating behavioral, economic, and technical insights to understand and address algorithmic bias: A human-centric perspective. *ACM Transactions on Management Information Systems*, *13*(3), 1–27 (cit. on pp. 75, 267).

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black box fairness testing of machine learning models. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 625–635 (cit. on pp. 68, 266, 267).

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *International Conference on Database Theory*, 420–434 (cit. on p. 88).

Aghaei, S., Azizi, M. J., & Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 1418–1426 (cit. on p. 267).

Ahn, Y., & Lin, Y.-R. (2020). FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1086–1095 (cit. on pp. 62, 63, 69, 70, 266, 267).

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., et al. (2019). Fairwashing: The risk of rationalization. *International Conference on Machine Learning*, 161–170 (cit. on pp. 67, 74, 266, 267).

Aïvodji, U., Ferry, J., Gambs, S., Huguet, M.-J., & Siala, M. (2021). FairCORELS, an open-source library for learning fair rule lists. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4665–4669 (cit. on pp. 71, 74, 266, 267).

Alexander, L., & Moore, M. (2020). Deontological ethics [https://plato.stanford.edu/entries/ethics-deontological/ (Accessed: 2023-07-08)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 212).

Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483* (cit. on pp. 68, 69, 266, 267).

Alufaisan, Y., Kantarcioglu, M., & Zhou, Y. (2021). Robust transparency against model inversion attacks. *IEEE Transactions on Dependable and Secure Computing*, *18*(5), 2061–2073 (cit. on pp. 65, 266).

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6618–6626 (cit. on p. 171).

Alves, G., Amblard, M., Bernier, F., Couceiro, M., & Napoli, A. (2021). Reducing unintended bias of ML models on tabular and textual data. *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10 (cit. on p. 267).

Alves, G., Bhargava, V., Bernier, F., Couceiro, M., & Napoli, A. (2020). FixOut: An ensemble approach to fairer models. *hal-03033181* (cit. on pp. 266, 267).

Alves, G., Bhargava, V., Couceiro, M., & Napoli, A. (2021). Making ML models fairer through explanations: The case of LimeOut. *AIST 2020: Analysis of Images, Social Networks and Texts*, 3–18 (cit. on pp. 68, 266, 267).

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989 (cit. on p. 35).

Anders, C., Pasliev, P., Dombrowski, A.-K., Müller, K.-R., & Kessel, P. (2020). Fairwashing explanations with off-manifold detergent. *International Conference on Machine Learning*, 314–323 (cit. on pp. 74, 267).

Anders, C., Weber, L., Neumann, D., Samek, W., Müller, K.-R., et al. (2022). Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *Information Fusion*, *77*, 261–295 (cit. on pp. 68, 266).

Anderson, E. S. (1999). What is the point of equality? *Ethics*, *109*(2), 287–337 (cit. on p. 40).

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction*, *4*(2), 556–579 (cit. on pp. 54, 67, 173, 266).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: 2023-04-10)]. *ProPublica*. (Cit. on pp. 4, 40, 55, 82, 108, 136).

Anik, A. I., & Bunt, A. (2021). Data-centric explanations: Explaining training data of machine learning systems to promote transparency. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 266).

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In blind pursuit of racial equality? *Psychological Science*, *21*(11), 1587–1592 (cit. on pp. 14, 169, 171, 174, 192).

Apple. (2022). How your Apple Card application is evaluated [https://support.apple.com/en-us/HT209218 (Accessed: 2023-05-06)]. *Apple*. (Cit. on p. 7).

Arneson, R. (2013). Egalitarianism [https://plato.stanford.edu/entries/egalitarianism/ (Accessed: 2023-06-10)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 39).

Arneson, R. (2015). Equality of opportunity [https://plato.stanford.edu/entries/equal-opportunity/ (Accessed: 2023-06-10)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 39).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115 (cit. on pp. 7, 9, 13, 26, 28, 30, 56, 58, 63, 64, 69, 109, 115, 136, 171, 173, 266, 267).

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., et al. (2021). AI Explainability 360 toolkit. *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 376–379 (cit. on pp. 8, 107).

Asher, N., de Lara, L., Paul, S., & Russell, C. (2022). Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, *4*(2), 316–349 (cit. on p. 266).

Ashktorab, Z., Desmond, M., Andres, J., Muller, M., Joshi, N. N., et al. (2021). AI-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–27 (cit. on p. 152).

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., et al. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1194–1206 (cit. on pp. 74, 267).

Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. (2022). Challenges in applying explainability methods to improve the fairness of NLP models. *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 80–92 (cit. on pp. 10, 56, 57, 266).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., et al. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16 (cit. on pp. 150, 151, 172).

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., et al. (2021). It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498* (cit. on p. 6).

Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). Studying up: Reorienting the study of algorithmic fairness around issues of power. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 167–176 (cit. on p. 51).

Barclay, D., Thompson, R., & Higgins, C. (1995). The partial least squares (PLS) approach to causal modeling: Personal computer use as an illustration. *Technology Studies*, *2*(2), 285–309 (cit. on pp. 120, 143).

Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., et al. (2021). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, *28*(3), 549–558 (cit. on p. 35).

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities* [http://www.fairmlbook.org (Accessed: 2023-06-19)]. (Cit. on pp. 7, 20, 23, 36, 41, 45–47, 50, 68, 173, 181, 206).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732 (cit. on pp. 42, 45, 55).

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, *143*(1), 30–56 (cit. on pp. 23, 106, 172).

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science*, *67*(5), 2964–2984 (cit. on p. 43).

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 1–21 (cit. on p. 109).

Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, *35*(2), 180–191 (cit. on pp. 117, 134).

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (cit. on pp. 69, 74, 266, 267).

Bélanger, F., Hiller, J. S., & Smith, W. J. (2002). Trustworthiness in electronic commerce: The role of privacy, security, and site attributes. *The Journal of Strategic Information Systems*, *11*(3–4), 245–270 (cit. on pp. 108, 139).

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., et al. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 400–422 (cit. on pp. 5, 6, 9, 48).

Bell, A., Stoyanovich, J., & Nov, O. (2023). Algorithmic transparency playbook [https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf (Accessed: 2023-05-10)]. *Center for Responsible AI*. (Cit. on p. 30).

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15 (cit. on pp. 8, 68).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (cit. on p. 42).

Berscheid, J., & Roewer-Despres, F. (2019). Beyond transparency: A proposed framework for accountability in decision-making AI systems. *AI Matters*, *5*(2), 13–22 (cit. on p. 266).

Bertsimas, D., Farias, V. F., & Trichakis, N. (2011). The price of fairness. *Operations Research*, *59*(1), 17–31 (cit. on p. 205).

Bhargava, V., Couceiro, M., & Napoli, A. (2020). LimeOut: An ensemble approach to improve process fairness. *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020)*, 475–491 (cit. on p. 267).

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., et al. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413 (cit. on p. 267).

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., et al. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657 (cit. on pp. 10, 170, 209).

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 405–414 (cit. on p. 84).

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Conference on Fairness, Accountability and Transparency*, 149–159 (cit. on pp. 39–41, 45).

Binns, R. (2022). Human judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197–211 (cit. on p. 23).

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524 (cit. on p. 48).

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., et al. (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*, 1–14 (cit. on pp. 32, 33, 52, 54, 63, 66, 67, 109, 111, 114, 116, 117, 133, 137, 173, 183, 266).

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI [https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf (Accessed: 2023-06-27)]. *Microsoft*. (Cit. on p. 8).

Black, E., Yeom, S., & Fredrikson, M. (2020). FlipTest: Fairness testing via optimal transport. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121 (cit. on pp. 266, 267).

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476 (cit. on pp. 57, 58).

Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias [https://apo.org.au/sites/default/files/resource-files/2018-12/apo-nid210071.pdf (Accessed: 2023-07-04)]. *Upturn*. (Cit. on p. 174).

Bommasani, R., Klyman, K., Zhang, D., & Liang, P. (2023). Do foundation model providers comply with the draft EU AI Act? [https://crfm.stanford.edu/2023/06/15/eu-ai-act.html?sf179246824=1 (Accessed: 2023-07-02)]. *Stanford University*. (Cit. on p. 203).

Bonham, V. L., Callier, S. L., & Royal, C. D. (2016). Will precision medicine move us beyond race? *The New England Journal of Medicine*, *374*(21), 2003–2005 (cit. on p. 51).

Borrellas, P., & Unceta, I. (2021). The challenges of machine learning and their economic implications. *Entropy*, *23*(3), 1–23 (cit. on pp. 75, 266, 267).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32 (cit. on pp. 89, 115, 117, 137, 141, 142).

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12 (cit. on pp. 111, 112).

Brown, T. (2008). Design thinking [https://hbr.org/2008/06/design-thinking (Accessed: 2023-07-08)]. *Harvard Business Review*. (Cit. on p. 214).

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258 (cit. on p. 123).

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*, 273–291 (cit. on p. 23).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (cit. on pp. 31, 203, 216).

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), 1–21 (cit. on pp. 152, 213).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91 (cit. on pp. 43, 108, 136, 138).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12 (cit. on p. 56).

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *2015 International Conference on Healthcare Informatics*, 160–169 (cit. on pp. 109, 172).

Buyl, M., Cociancig, C., Frattone, C., & Roekens, N. (2022). Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1071–1082 (cit. on p. 172).

Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving human-AI collaboration with descriptions of AI behavior. *arXiv preprint arXiv:2301.06937* (cit. on p. 152).

Cai, L., & Liu, X. (2022). Identifying Big Five personality traits based on facial behavior analysis. *Frontiers in Public Health*, *10*, 1001828 (cit. on p. 44).

Calegari, R., Ciatto, G., Denti, E., & Omicini, A. (2020). Logic-based technologies for intelligent systems: State of the art and perspectives. *Information*, *11*(3), 167 (cit. on p. 266).

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, *14*(1), 7–32 (cit. on p. 266).

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, *30* (cit. on p. 84).

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). AI Now 2017 Report [https://ainowinstitute.org/publication/ai-now-2017-report-2 (Accessed: 2023-06-19)]. *AI Now Institute at New York University*. (Cit. on p. 36).

Candelon, F., Evgeniou, T., & Martens, D. (2023). AI can be both accurate and transparent [https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent? (Accessed: 2023-05-14)]. *Harvard Business Review*. (Cit. on pp. 10, 31).

Carey, D., & Smith, M. (2016). How companies are using simulations, competitions, and analytics to hire [https://hbr.org/2016/04/how-companies-are-using-simulations-competitions-and-analytics-to-hire (Accessed: 2023-06-17)]. *Harvard Business Review*. (Cit. on p. 108).

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information Systems Journal*, *15*(1), 5–25 (cit. on pp. 116, 141).

Castelluccia, C., & Le Métayer, D. (2019). Understanding algorithmic decision-making: Opportunities and challenges [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf (Accessed: 2023-04-10)]. *European Parliamentary Research Service*. (Cit. on p. 81).

Castelnovo, A., Malandri, L., Mercorio, F., Mezzanzanica, M., & Cosentini, A. (2021). Towards fairness through time. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 647–663 (cit. on pp. 68, 266).

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825 (cit. on pp. 138, 143).

Castillo, C. (2019). Fairness and transparency in ranking. *ACM SIGIR Forum*, *52*(2), 64–71 (cit. on p. 84).

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080 (cit. on pp. 56, 65, 67, 266).

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (cit. on p. 9).

Celis, L. E., Straszak, D., & Vishnoi, N. K. (2018). Ranking with fairness constraints. *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, 28:1–28:15 (cit. on p. 84).

Cesaro, J., & Gagliardi Cozman, F. (2019). Measuring unfairness through game-theoretic interpretability. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 253–264 (cit. on pp. 68, 266).

Chakraborty, J., Peng, K., & Menzies, T. (2020). Making fair ML software using trustworthy explanation. *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, 1229–1233 (cit. on pp. 71, 74, 75, 267).

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., et al. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–127 (cit. on pp. 22, 106, 108).

Chan, D. (2011). Perceptions of fairness. *Research Collection School of Social Sciences, Singapore Management University*, 2796 (cit. on pp. 52, 108, 139).

Chan, G. K. Y. (2022). AI employment decision-making: Integrating the equal opportunity merit principle and explainable AI. *AI & Society*, 1–12 (cit. on pp. 73, 266, 267).

Charmaz, K. (2003). Grounded theory. In *Qualitative psychology: A practical guide to research methods* (pp. 81–110). Sage Publications, Inc. (Cit. on p. 125).

Chatterjee, D. (2019). Loan prediction problem dataset [https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset (Accessed: 2021-08-24)]. *Kaggle*. (Cit. on pp. 114, 141).

Chaudoin, S., Gaines, B. J., & Livny, A. (2021). Survey design, order effects, and causal mediation analysis. *The Journal of Politics*, *83*(4), 1851–1856 (cit. on p. 176).

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*, 1–23 (cit. on p. 55).

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255* (cit. on p. 170).

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern Methods for Business Research*, *295*(2), 295–336 (cit. on p. 120).

Chiu, C.-M., Lin, H.-Y., Sun, S.-Y., & Hsu, M.-H. (2009). Understanding customers' loyalty intentions towards online shopping: An integration of technology acceptance model and fairness theory. *Behaviour & Information Technology*, *28*(4), 347–360 (cit. on pp. 116, 141).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163 (cit. on pp. 5, 6, 9, 46–48, 76, 106, 115, 173, 181).

Chromik, M., Eiband, M., Völkel, S. T., & Buschek, D. (2019). Dark patterns of explainability, transparency, and user control for intelligent systems. *Joint Proceedings of the ACM IUI 2019 Workshops* (cit. on pp. 9, 27, 34, 109, 170, 207).

Chun Tie, Y., Birks, M., & Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, *7*, 1–8 (cit. on p. 61).

Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., & Whang, S. E. (2019). Automated data slicing for model validation: A big data-AI integration approach. *IEEE Transactions on Knowledge and Data Engineering*, *32*(12), 2284–2296 (cit. on pp. 68, 266).

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, *99*(4), 906–944 (cit. on p. 39).

Colaner, N. (2022). Is explainable artificial intelligence intrinsically valuable? *AI & Society*, *37*, 1–8 (cit. on p. 266).

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *The Journal of Applied Psychology*, *86*(3), 386–400 (cit. on p. 66).

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425–445 (cit. on pp. 52, 141).

Colquitt, J. A., & Rodell, J. B. (2011). Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives. *Academy of Management Journal*, *54*(6), 1183–1206 (cit. on pp. 112, 113).

Colquitt, J. A., & Rodell, J. B. (2015). Measuring justice and fairness. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice in the workplace* (pp. 187–202). Oxford University Press. (Cit. on pp. 52, 107, 108, 112, 116, 173, 182, 183).

Colson, E. (2019). What AI-driven decision making looks like [https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like (Accessed: 2023-05-19)]. *Harvard Business Review*. (Cit. on pp. 19, 22).

Cook, K. S., & Hegtvedt, K. A. (1983). Distributive justice, equity, and equality. *Annual Review of Sociology*, *9*(1), 217–241 (cit. on p. 6).

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (cit. on pp. 9, 43, 54, 169, 171, 173, 174).

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806 (cit. on pp. 84, 99).

Cornacchia, G., Narducci, F., & Ragone, A. (2021). A general model for fair and explainable recommendation in the loan domain. *Proceedings of the Joint KaRS & ComplexRec Workshop* (cit. on pp. 71, 266, 267).

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104 (cit. on pp. 120, 143).

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313 (cit. on p. 138).

Culyer, A. J., & Wagstaff, A. (1993). Equity and equality in health and health care. *Journal of Health Economics*, *12*(4), 431–457 (cit. on p. 6).

Dai, J., Upadhyay, S., Aïvodji, U., Bach, S. H., & Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 203–214 (cit. on pp. 74, 267).

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691–4697 (cit. on pp. 41, 42).

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371* (cit. on p. 173).

Dash, S., Balasubramanian, V. N., & Sharma, A. (2022). Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 915–924 (cit. on pp. 70, 266, 267).

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296–299). Auerbach Publications. (Cit. on pp. 6, 44).

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (cit. on p. 35).

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy*, 598–617 (cit. on pp. 63, 69, 266, 267).

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6), 100489 (cit. on p. 26).

De-Arteaga, M., Dubrawski, A., & Chouldechova, A. (2018). Learning under selective labels in the presence of expert consistency. *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)* (cit. on p. 21).

De-Arteaga, M., Feuerriegel, S., & Saar-Tsechansky, M. (2022). Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, *31*(10), 3749–3770 (cit. on pp. 36, 41–44, 46–49, 51, 152, 172, 205, 211).

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12 (cit. on pp. 24, 25, 109, 134, 172, 192).

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., et al. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128 (cit. on pp. 168, 174, 175, 180, 187).

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, *39*(2), 101666 (cit. on p. 170).

Decety, J., & Fotopoulou, A. (2015). Why empathy has a beneficial impact on others in medicine: Unifying theories. *Frontiers in Behavioral Neuroscience*, *8*, 457 (cit. on p. 23).

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society*, *35*, 917–926 (cit. on p. 267).

de Greeff, J., de Boer, M. H., Hillerström, F. H., Bomhof, F., Jorritsma, W., et al. (2021). The FATE system: Fair, transparent and explainable decision making. *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering* (cit. on pp. 266, 267).

de Vogue, A., Cole, D., & Sneed, T. (2023). Supreme Court guts affirmative action in college admissions [https://www.cnn.com/2023/06/29/politics/affirmative-action-supreme-court-ruling/index.html (Accessed: 2023-07-01)]. *CNN*. (Cit. on p. 40).

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62 (cit. on p. 35).

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity [http://go.volarisgroup.com/rs/430-MBX-989/images/%20ProPublica_Commentary_Final_070616.pdf (Accessed: 2023-06-27)]. *Northpointe Inc. Research Department*. (Cit. on p. 5).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126 (cit. on pp. 25, 138, 139).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170 (cit. on p. 112).

Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. *SafeAI@AAAI* (cit. on pp. 34, 74, 171, 206, 209, 266, 267).

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285 (cit. on pp. 10, 13, 54, 63, 67, 109, 137, 168, 170, 171, 173, 266, 267).

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818 (cit. on p. 42).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (cit. on pp. 30, 57, 63, 266).

Du, M., Yang, F., Zou, N., & Hu, X. (2021). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, *36*(4), 25–34 (cit. on pp. 266, 267).

Dua, D., & Graff, C. (2017). UCI machine learning repository [http://archive.ics.uci.edu/ml (Accessed: 2023-04-10)]. *University of California, Irvine*. (Cit. on p. 96).

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335 (cit. on p. 42).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226 (cit. on pp. 12, 37, 48, 50, 51, 83, 84, 86, 93, 103, 169, 171, 174, 192, 198, 209).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718 (cit. on p. 172).

Edelman. (2021). 2021 Edelman Trust Barometer: Trust in technology [https://www.edelman.com/trust/2021-trust-barometer/trust-technology (Accessed: 2023-06-17)]. *Edelman*. (Cit. on pp. 106, 138, 139).

Ehsan, U., & Riedl, M. O. (2021). Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (cit. on pp. 35, 67, 109, 171).

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. *International Conference on Human-Computer Interaction*, 449–466 (cit. on pp. 41, 170).

Ehsan, U., Saha, K., De Choudhury, M., & Riedl, M. O. (2023). Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 34:1–34:32 (cit. on pp. 213, 214).

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6 (cit. on pp. 9, 34, 67, 171).

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650 (cit. on pp. 32, 170, 209).

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., et al. (2015). "I always assumed that I wasn't really that close to [her]" – Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162 (cit. on p. 138).

Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., et al. (2019). User attitudes towards algorithmic opacity and transparency in online reviewing platforms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on p. 109).

ETS. (2019). GRE guide to the use of scores 2019–20 [https://www.ets.org/s/gre/pdf/gre_guide.pdf (Accessed: 2020-04-23)]. *Educational Testing Service*. (Cit. on pp. 86, 99).

ETS. (2018). A snapshot of the individuals who took the GRE General Test [https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2018.pdf (Accessed: 2020-02-05)]. *Educational Testing Service*. (Cit. on pp. 86, 99).

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. (Cit. on p. 22).

European Union. (2016). General data protection regulation [https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed: 2023-06-17)]. *European Union*. (Cit. on pp. 8, 106, 170, 192).

Fabris, A., Mishler, A., Gottardi, S., Carletti, M., Daicampi, M., et al. (2021). Algorithmic audit of Italian car insurance: Evidence of unfairness in access and pricing. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 458–468 (cit. on p. 23).

Fan, M., Wei, W., Jin, W., Yang, Z., & Liu, T. (2022). Explanation-guided fairness testing through genetic algorithm. *Proceedings of the 44th International Conference on Software Engineering*, 871–882 (cit. on pp. 266, 267).

Federal Anti-Discrimination Agency. (2006). General equal treatment act [https://www.antidiskriminierungsstelle.de/EN/about-discrimination/order-and-law/general-equal-treatment-act/general-equal-treatment-act-node (Accessed: 2023-06-24)]. *Federal Anti-Discrimination Agency (FADA)*. (Cit. on p. 41).

Fernandez, C., Provost, F., & Han, X. (2019). Counterfactual explanations for data-driven decisions. *ICIS 2019 Proceedings*, 8 (cit. on p. 137).

Ferreira, J. J., & Monteiro, M. d. S. (2020). Evidence-based explanation to promote fairness in AI systems. *arXiv preprint arXiv:2003.01525* (cit. on pp. 64, 65, 204, 266).

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and opportunities. *Business & Information Systems Engineering*, *62*, 379–384 (cit. on p. 106).

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks". *Federal Probation*, *80*(2), 38–46 (cit. on pp. 5, 6).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., et al. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 689–707 (cit. on pp. 56, 63, 65, 72, 266, 267).

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, *26*, 1771–1796 (cit. on p. 267).

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50 (cit. on pp. 120, 143).

Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing*, *470*, 318–334 (cit. on p. 266).

Franke, U. (2022). First- and second-level bias in automated decision-making. *Philosophy & Technology*, *35*(21), 1–20 (cit. on pp. 70, 266, 267).

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236* (cit. on pp. 5, 9).

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: Testing software for discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510 (cit. on p. 266).

Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. *Proceedings of the 2021 International Conference on Management of Data*, 577–590 (cit. on pp. 68, 266, 267).

Galinkin, E. (2022). Towards a responsible AI development lifecycle: Lessons from information security. *arXiv preprint arXiv:2203.02958* (cit. on pp. 74, 266, 267).

Ganju, K. K., Atasoy, H., McCullough, J., & Greenwood, B. (2020). The role of decision support systems in attenuating racial biases in healthcare delivery. *Management Science*, *66*(11), 5171–5181 (cit. on p. 211).

Garcia, M. A. R., Rojas, R., Gualtieri, L., Rauch, E., & Matt, D. (2019). A human-in-the-loop cyber-physical system for collaborative assembly in smart manufacturing. *Procedia CIRP*, *81*, 600–605 (cit. on p. 24).

Ge, Y., Tan, J., Zhu, Y., Xia, Y., Luo, J., et al. (2022). Explainable fairness in recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 681–691 (cit. on pp. 70, 267).

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., et al. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92 (cit. on pp. 49, 214).

Ghosh, A., Shanbhag, A., & Wilson, C. (2022). FairCanary: Rapid continuous explainable fairness. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 307–316 (cit. on pp. 68, 70, 267).

Ghosh, B., Basu, D., & Meel, K. S. (2023). "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 138–148 (cit. on pp. 69, 267).

Gill, N., Hall, P., Montgomery, K., & Schmidt, N. (2020). A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing. *Information*, *11*(3), 1–32 (cit. on pp. 56, 63, 74, 75, 266, 267).

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., et al. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89 (cit. on pp. 10, 13, 64, 67, 73, 170, 171, 209, 266, 267).

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127 (cit. on pp. 25, 145).

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375 (cit. on pp. 109, 134).

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., et al. (2018). Explainable AI: The new 42? *2nd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, 295–303 (cit. on pp. 109, 137).

Gol Mohammadi, N., Paulus, S., Bishr, M., Metzger, A., Könnecke, H., et al. (2013). Trustworthiness attributes and metrics for engineering trusted internet-based software systems. *International Conference on Cloud Computing and Services Science*, 19–35 (cit. on p. 53).

Goldman, B., & Cropanzano, R. (2015). "Justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, *36*(2), 313–318 (cit. on pp. 37, 38).

Goodin, R. E. (1999). Treating likes alike, intergenerationally and internationally. *Policy Sciences*, *32*(2), 189–206 (cit. on p. 37).

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57 (cit. on pp. 8, 10, 67, 106, 210).

Google. (2023). Responsible AI practices [https://ai.google/responsibilities/responsible-ai-practices/ (Accessed: 2023-05-06)]. *Google*. (Cit. on p. 8).

Gorny, P. M., Nieken, P., & Ströhlein, K. (2023a). The effects of gendered language on norm compliance. *CESifo Working Papers*, 10459 (cit. on p. 175).

Gorny, P. M., Nieken, P., & Ströhlein, K. (2023b). He, she, they? The impact of gendered language on economic behavior. *CESifo Working Papers*, 10458 (cit. on p. 175).

Gosepath, S. (2021). Equality [https://plato.stanford.edu/entries/equality/ (Accessed: 2023-06-30)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 37).

Grabowicz, P. A., Perello, N., & Mishra, A. (2022). Marrying fairness and explainability in supervised learning. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1905–1916 (cit. on pp. 63, 69, 266, 267).

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on p. 207).

Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99 (cit. on pp. 24, 25).

Green, B., & Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24 (cit. on p. 171).

Greenberg, J. (1987). A taxonomy of organizational justice theories. *Academy of Management Review*, *12*(1), 9–22 (cit. on pp. 52, 173).

Grgić-Hlača, N., Lima, G., Weller, A., & Redmiles, E. M. (2022). Dimensions of diversity in human perceptions of algorithmic fairness. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–12 (cit. on pp. 54, 110).

Grgić-Hlača, 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*, 903–912 (cit. on pp. 54, 110, 173).

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1) (cit. on pp. 39, 50, 54, 107, 110, 173).

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2020). 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 (NIPS 2016)* (cit. on pp. 9, 54, 83, 84, 173, 174).

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, *83*(2), 241–262 (cit. on p. 267).

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211 (cit. on pp. 22, 106, 108, 136).

Gryz, J., & Shahbazi, N. (2020). Futility of a right to explanation. *EDBT/ICDT Workshops* (cit. on pp. 72, 75, 267).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42 (cit. on pp. 32, 109, 170).

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58 (cit. on p. 10).

Gupta, V., Nokhiz, P., Roy, C. D., & Venkatasubramanian, S. (2019). Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166* (cit. on pp. 67, 70, 266, 267).

Hacker, P., & Passoth, J.-H. (2020). Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 343–373 (cit. on pp. 266, 267).

Haddaway, N. R., Grainger, M. J., & Gray, C. T. (2022). Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods*, *13*(4), 533–545 (cit. on p. 60).

Hair Jr., J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage Publications. (Cit. on p. 123).

Hall, P., & Gill, N. (2017). Debugging the black-box COMPAS risk assessment instrument to diagnose and remediate bias. *OpenReview.net* (cit. on pp. 266, 267).

Hall, P., Gill, N., & Schmidt, N. (2019). Proposed guidelines for the responsible use of explainable machine learning. *arXiv preprint arXiv:1906.03533* (cit. on p. 267).

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 44).

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, *17*(1), 72–85 (cit. on p. 267).

Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., et al. (2021). Amazon SageMaker clarify: Machine learning bias detection and explainability in the cloud. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2974–2983 (cit. on pp. 63, 266).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29* (cit. on pp. 46, 50, 83, 84, 89, 173, 182).

Harris, J. G., & Davenport, T. H. (2005). Automated decision making comes of age [https://sloanreview.mit.edu/article/automated-decision-making-comes-of-age/ (Accessed: 2023-06-19)]. *MIT Sloan Management Review*. (Cit. on pp. 3, 22, 106, 108, 136).

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*, 392–402 (cit. on pp. 9, 54, 174).

Hayes, P. (2020). An ethical intuitionist account of transparency of algorithms and its gradations. *Business Research*, *13*(3), 849–874 (cit. on p. 212).

He, Y., Burghardt, K., & Lerman, K. (2020). A geometric solution to fair representations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 279–285 (cit. on p. 267).

Heaven, W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled [https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/ (Accessed: 2023-06-19)]. *MIT Technology Review*. (Cit. on pp. 3, 106, 108, 136).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161 (cit. on p. 36).

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2022). On the effect of information asymmetry in human-AI teams. *ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI)* (cit. on pp. 193, 213).

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. *PACIS 2021 Proceedings*, 78 (cit. on pp. 24, 25, 151).

Hemmer, P., Thede, L., Vössing, M., Jakubik, J., & Kühl, N. (2023). Learning to defer with limited expert predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(5), 6002–6011 (cit. on pp. 25, 215).

Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414* (cit. on pp. 73, 74, 267).

Hickey, J. M., Di Stefano, P. G., & Vasileiou, V. (2021). Fairness by explicability and adversarial SHAP learning. *ECML PKDD 2020: Machine Learning and Knowledge Discovery in Databases*, 174–190 (cit. on pp. 70, 72, 267).

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170355 (cit. on p. 23).

Hind, M., Wei, D., Campbell, M., Codella, N. C., Dhurandhar, A., et al. (2019). TED: Teaching AI to explain its decisions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129 (cit. on pp. 64, 67, 71, 266, 267).

Hitlin, P. (2016). Research in the crowdsourcing age: A case study [https://www.pewinternet.org/wp-content/uploads/sites/9/2016/07/PI_2016.07.11_Mechanical-Turk_FINAL.pdf (Accessed: 2023-07-03)]. *Pew Research Center*. (Cit. on p. 110).

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). GAMUT: A design probe to understand how data scientists understand machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 266).

Holstein, K., De-Arteaga, M., Tumati, L., & Cheng, Y. (2023). Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–20 (cit. on pp. 193, 213).

Holzer, H., & Neumark, D. (2000). Assessing affirmative action. *Journal of Economic Literature*, *38*(3), 483–568 (cit. on pp. 40, 133).

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058* (cit. on pp. 41, 42, 44).

Houlden, P., LaTour, S., Walker, L., & Thibaut, J. (1978). Preference for modes of dispute resolution as a function of process and decision control. *Journal of Experimental Social Psychology*, *14*(1), 13–30 (cit. on pp. 112, 131).

Hu, K. (2023). ChatGPT sets record for fastest-growing user base: Analyst note [https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Accessed: 2023-05-20)]. *Reuters*. (Cit. on pp. 22, 216).

Hu, L., Immorlica, N., & Wortman Vaughan, J. (2019). The disparate effects of strategic manipulation. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268 (cit. on p. 43).

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 513–513 (cit. on pp. 48, 84).

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58 (cit. on p. 41).

IBM. (2023). AI ethics [https://www.ibm.com/artificial-intelligence/ethics (Accessed: 2023-05-06)]. *IBM*. (Cit. on p. 8).

Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E., & Marques-Silva, J. (2020). Towards formal fairness in machine learning. *International Conference on Principles and Practice of Constraint Programming*, 846–867 (cit. on p. 266).

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. *Proceedings of the Web Conference 2021*, 3767–3778 (cit. on pp. 23, 106, 172).

Infosys. (2019). How FinTechs can enable better support to FIs' credit decisioning? [https://www.infosys.com/industries/financial-services/insights/documents/fintechs-fi-partners-credit-decision.pdf (Accessed: 2023-06-17)]. *Infosys*. (Cit. on pp. 114, 115, 141).

Ingold, D., & Soper, S. (2016). Amazon doesn't consider the race of its customers. Should it? [https://www.bloomberg.com/graphics/2016-amazon-same-day/ (Accessed: 2023-06-18)]. *Bloomberg*. (Cit. on p. 51).

Jabbari, S., Ou, H.-C., Lakkaraju, H., & Tambe, M. (2020). An empirical study of the trade-offs between interpretability and fairness. *ICML Workshop on Human Interpretability in Machine Learning*, 1–6 (cit. on pp. 75, 267).

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205 (cit. on p. 33).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635 (cit. on pp. 53, 108, 172).

Jain, A., Ravula, M., & Ghosh, J. (2020). Biased models have biased explanations. *arXiv preprint arXiv:2012.10986* (cit. on pp. 68, 266, 267).

Jakubik, J., Schöffer, J., Hoge, V., Vössing, M., & Kühl, N. (2023). An empirical evaluation of predicted outcomes as explanations in human-AI decision-making. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022*, 353–368 (cit. on p. 152).

Jernite, Y. (2022). Machine learning in development: Let's talk about bias! [https://huggingface.co/blog/ethics-soc-2 (Accessed: 2023-06-24)]. *Hugging Face*. (Cit. on pp. 41, 42).

Jillson, E. (2021). Aiming for truth, fairness, and equity in your company's use of AI [https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai (Accessed: 2023-06-19)]. *Federal Trade Commission*. (Cit. on p. 29).

John-Mathews, J.-M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change*, *174*, 121209 (cit. on pp. 63, 67, 72, 266, 267).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260 (cit. on p. 22).

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, *29* (cit. on pp. 53, 110).

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589 (cit. on p. 31).

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6 (cit. on p. 84).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33 (cit. on pp. 49, 84).

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*, 924–929 (cit. on p. 50).

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50 (cit. on pp. 50, 84).

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, *55*(5), 1–29 (cit. on pp. 70, 266, 267).

Kasirzadeh, A., & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 228–236 (cit. on p. 267).

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., et al. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175 (cit. on p. 110).

Kearns, M., Roth, A., & Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. *International Conference on Machine Learning*, 1828–1836 (cit. on p. 85).

Kehrenberg, T., Bartlett, M., Thomas, O., & Quadrianto, N. (2020). Null-sampling for interpretable and fair representations. *ECCV 2020: European Conference on Computer Vision*, 565–580 (cit. on p. 267).

Kelion, L. (2019). Apple's 'sexist' credit card investigated by US regulator [https://www.bbc.com/news/business-50365609 (Accessed: 2023-05-06)]. *British Broadcasting Corporation (BBC)*. (Cit. on p. 7).

Kenny, D. A. (2015). Measuring model fit [http://www.davidakenny.net/cm/fit.htm (Accessed: 2023-07-03)]. (Cit. on pp. 123, 131).

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., & Valera, I. (2020). Fair decisions despite imperfect predictions. *International Conference on Artificial Intelligence and Statistics*, 277–287 (cit. on pp. 82, 84).

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., et al. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, *30* (cit. on p. 84).

Kim, A., Yang, M., & Zhang, J. (2023). When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction*, *30*(1), 1–36 (cit. on pp. 152, 180).

Kitchenham, B., & Charters, S. M. (2007). Guidelines for performing systematic literature reviews in software engineering [https://cdn.elsevier.com/promis_misc/525444systematicreviewsguide.pdf (Accessed: 2023-07-01)]. *Keele University, University of Durham*. (Cit. on pp. 59, 60).

Kite-Powell, J. (2022). Explainable AI is trending and here's why [https://www.forbes.com/sites/jenniferhicks/2022/07/28/explainable-ai-is--trending-and-heres-why/ (Accessed: 2023-06-19)]. *Forbes*. (Cit. on pp. 10, 168, 209).

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395 (cit. on pp. 110, 138, 171, 266).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, 22–27 (cit. on pp. 14, 39, 50, 169, 171, 174, 192).

Kleinberg, J., & Mullainathan, S. (2019). Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 807–808 (cit. on pp. 11, 42, 44, 63, 75, 205, 267).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43:1–43:23 (cit. on pp. 5, 9, 48, 76, 106, 115).

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications. (Cit. on p. 123).

Koivunen, S., Olsson, T., Olshannikova, E., & Lindberg, A. (2019). Understanding decision-making in recruitment: Opportunities and challenges for information technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(GROUP), 1–22 (cit. on pp. 22, 106, 108).

Kop, M. (2021). EU artificial intelligence act: The European approach to AI. *Transatlantic Antitrust and IPR Developments* (cit. on p. 29).

Kramer, M. F., Schaich Borg, J., Conitzer, V., & Sinnott-Armstrong, W. (2018). When do people want AI to make decisions? *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 204–209 (cit. on p. 138).

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(633), 633–705 (cit. on pp. 63, 266, 267).

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621 (cit. on pp. 121, 184).

Kuckartz, U., & Rädiker, S. (2019). *Analyzing qualitative data with MAXQDA: Text, audio, and video*. Springer. (Cit. on pp. 61, 62, 125).

Kühl, N., Lobana, J., & Meske, C. (2020). Do you comply with AI? – Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *ICIS 2019 Proceedings*, 1 (cit. on p. 132).

Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets*, *32*, 2235–2244 (cit. on pp. 20, 21).

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., et al. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10 (cit. on p. 109).

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct [https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct (Accessed: 2023-06-19)]. *Harvard Business Review*. (Cit. on pp. 3, 22, 105, 106, 108, 136).

Kung, C., & Yu, R. (2020). Interpretable models do not compromise accuracy or fairness in predicting college success. *L@S '20: Proceedings of the Seventh ACM Conference on Learning @ Scale*, 413–416 (cit. on pp. 63, 267).

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30* (cit. on pp. 36, 48, 50, 84, 169).

Laato, S., Tiainen, M., Najmul Islam, A., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31 (cit. on p. 56).

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471* (cit. on pp. 7, 24, 25, 53, 150–152, 162, 172, 173, 207).

Lai, V., Liu, H., & Tan, C. (2020). "Why is 'Chicago' deceptive?" Towards building model-driven tutorials for humans. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on pp. 33, 152, 177).

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38 (cit. on pp. 4, 22, 171).

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?" Manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85 (cit. on pp. 9, 34, 171, 192, 206).

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284 (cit. on pp. 21, 82, 84).

Lamont, J., & Favor, C. (2017). Distributive justice [https://plato.stanford.edu/entries/justice-distributive/ (Accessed: 2023-06-16)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 38).

Lance Frazier, M., Johnson, P. D., Gavin, M., Gooty, J., & Bradley Snow, D. (2010). Organizational justice, trustworthiness, and trust: A multifoci examination. *Group & Organization Management*, *35*(1), 39–76 (cit. on p. 113).

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., et al. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, *29*(2), 154–169 (cit. on p. 109).

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, 106878 (cit. on pp. 64, 65).

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., et al. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473 (cit. on pp. 10, 25–30, 34, 56, 57, 63, 65, 68, 70, 76, 107, 109, 170, 173, 204, 212, 214, 261, 266, 267).

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm [https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Accessed: 2023-05-05)]. *ProPublica*. (Cit. on p. 5).

Lawless, C., Dash, S., Gunluk, O., & Wei, D. (2021). Interpretable and fair Boolean rule sets via column generation. *arXiv preprint arXiv:2111.08466* (cit. on p. 267).

Lazar, S. (2022). Legitimacy, authority, and the political value of explanations. *arXiv preprint arXiv:2208.08628* (cit. on pp. 29, 30, 204, 214).

Le Merrer, E., & Trédan, G. (2020). Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, *2*(9), 529–539 (cit. on pp. 266, 267).

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126–148 (cit. on pp. 193, 213).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444 (cit. on pp. 21, 31).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80 (cit. on pp. 25, 108, 172).

Lee, M. S. A. (2019). Context-conscious fairness in using machine learning to make decisions. *AI Matters*, *5*(2), 23–29 (cit. on p. 66).

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16 (cit. on pp. 104, 110, 116, 137–139, 141).

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048 (cit. on pp. 110, 138).

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26 (cit. on pp. 53, 63, 107, 108, 110, 112, 138, 173, 266).

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*, 1–14 (cit. on p. 110).

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., et al. (2022). Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (cit. on p. 20).

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., et al. (2022). Combining the strengths of radiologists and AI for breast cancer screening: A retrospective analysis. *The Lancet Digital Health*, *4*(7), e507–e519 (cit. on pp. 3, 24, 25, 215).

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611–627 (cit. on pp. 63, 267).

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? The bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data* (pp. 3–24). Springer. (Cit. on pp. 22, 106).

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector [https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf (Accessed: 2023-06-18)]. *The Alan Turing Institute*. (Cit. on pp. 64, 69, 72, 266, 267).

Lewis, C., & Mack, R. (1982). The role of abduction in learning to use a computer system. *Annual Meeting of the American Educational Research Association*, 1–11 (cit. on p. 109).

Liao, Q. V., & Wortman Vaughan, J. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (cit. on p. 34).

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2119–2128 (cit. on p. 107).

Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2103–2113 (cit. on p. 34).

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(18), 1–45 (cit. on p. 267).

Lind, E. A., Lissak, R. I., & Conlon, D. E. (1983). Decision control and process control effects on procedural fairness judgments. *Journal of Applied Social Psychology*, *13*(4), 338–350 (cit. on p. 112).

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57 (cit. on pp. 10, 57, 63, 68, 74, 170, 193, 207, 212, 213, 266, 267).

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–45 (cit. on pp. 171, 175, 177, 181).

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., & Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (cit. on p. 110).

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology*, *23*(3), 253–263 (cit. on pp. 72, 73, 266, 267).

Lombrozo, T. (2012). Explanation and abductive inference. In *The Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press. (Cit. on pp. 33, 212).

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16 (cit. on pp. 139–141).

Long, R. (2021). Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy*, *19*(1), 49–78 (cit. on pp. 107, 108).

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16 (cit. on pp. 151, 152).

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19 (cit. on p. 35).

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67 (cit. on pp. 32, 170).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30* (cit. on pp. 7, 32, 170, 209).

Lynch, S. (2023). 2023 state of AI in 14 charts [https://hai.stanford.edu/news/2023-state-ai-14-charts (Accessed: 2023-07-08)]. *Stanford University*. (Cit. on p. 216).

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psycholology*, *58*, 593–614 (cit. on p. 131).

Maclure, J. (2021). AI, explainability and public reason: The argument from the limitations of the human mind. *Minds and Machines*, *31*(3), 421–438 (cit. on p. 267).

Madiega, T. (2021). Artificial intelligence act [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf (Accessed: 2023-05-08)]. *European Parliamentary Research Service*. (Cit. on pp. 8, 23, 42, 203).

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, *31* (cit. on pp. 25, 215).

Manerba, M. M., & Guidotti, R. (2021). FairShades: Fairness auditing via explainability in abusive language detection systems. *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, 34–43 (cit. on pp. 266, 267).

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60 (cit. on pp. 121, 184).

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–130 (cit. on p. 183).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734 (cit. on p. 53).

Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal*, *128*(8), 2218–2300 (cit. on p. 174).

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, *27*(4), 12–12 (cit. on p. 19).

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160 (cit. on p. 42).

McFarland, M. (2016). Terrorist or pedophile? This start-up says it can out secrets by analyzing faces [https://www.washingtonpost.com/news/innovations/wp/2016/05/24/terrorist-or-pedophile-this-start-up-says-it-can-out-secrets-by-analyzing-faces/ (Accessed: 2023-06-25)]. *The Washington Post*. (Cit. on p. 44).

McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15458–15463 (cit. on p. 7).

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U test. *The Corsini Encyclopedia of Psychology* (cit. on p. 143).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35 (cit. on pp. 22, 36, 46, 49, 83, 100, 106).

Mencl, J., & May, D. R. (2009). The effects of proximity and empathy on ethical decision-making: An exploratory investigation. *Journal of Business Ethics*, *85*, 201–226 (cit. on p. 23).

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports*, *12*(7166), 1–28 (cit. on pp. 57, 68, 70, 266, 267).

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63 (cit. on pp. 136, 137).

Michael, L. (2019). Machine coaching. *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, 80–86 (cit. on p. 266).

Millea, M., Wills, R., Elder, A., & Molina, D. (2018). What matters in college student success? Determinants of college retention and graduation rates. *Education*, *138*(4), 309–322 (cit. on p. 27).

Miller, C. C. (2015). Can an algorithm hire better than a human? [https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html (Accessed: 2023-06-15)]. *The New York Times*. (Cit. on p. 41).

Miller, D. (2021). Justice [https://plato.stanford.edu/entries/justice/ (Accessed: 2023-06-16)]. *Stanford Encyclopedia of Philosophy*. (Cit. on pp. 37–40).

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28*(4), 283–298 (cit. on pp. 170, 175, 186).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38 (cit. on pp. 33, 109, 212).

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, *29*(2), 111–147 (cit. on pp. 63, 266, 267).

Miroshnikov, A., Kotsiopoulos, K., Franks, R., & Ravi Kannan, A. (2022). Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning*, *111*(9), 3307–3357 (cit. on pp. 69, 266, 267).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163 (cit. on p. 41).

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21 (cit. on pp. 65, 266).

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288 (cit. on p. 33).

Molnar, C. (2020). *Interpretable machine learning* [https://christophm.github.io/interpretable-ml-book/ (Accessed: 2023-06-19)]. (Cit. on pp. 7, 12, 21, 30–32, 35, 109, 198, 213).

Mourão, E., Pimentel, J. F., Murta, L., Kalinowski, M., Mendes, E., et al. (2020). On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and Software Technology*, *123*, 106294 (cit. on p. 59).

Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. *International Conference on Machine Learning*, 7076–7087 (cit. on pp. 25, 151, 215).

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–36 (cit. on pp. 9, 46, 115).

Mutlu, E. Ç., Yousefi, N., & Ozmen Garibay, O. (2022). Contrastive counterfactual fairness in algorithmic decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 499–507 (cit. on p. 267).

Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., & Strappelli, L. (2022). Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *12*(3), 1–30 (cit. on p. 266).

Narayanan, A., Mathur, A., Chetty, M., & Kshirsagar, M. (2020). Dark patterns: Past, present, and future. *Queue*, *18*(2), 67–92 (cit. on p. 34).

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., et al. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (cit. on p. 171).

Nasiripour, S., & Natarajan, S. (2019). Apple co-founder says Goldman's Apple Card algorithm discriminates [https://www.bloomberg.com/news/articles/2019-11-10/apple-co-founder-says-goldman-s-apple-card-algo-discriminates (Accessed: 2023-04-10)]. *Bloomberg*. (Cit. on p. 82).

Nassih, R., & Berrado, A. (2020). State of the art of fairness, interpretability and explainability in machine learning: Case of PRIM. *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 1–5 (cit. on p. 266).

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*, *24*(1), 3–14 (cit. on pp. 3, 22, 106).

Noon, M. (2010). The shackled runner: Time to rethink positive discrimination? *Work, Employment and Society*, *24*(4), 728–739 (cit. on p. 40).

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (cit. on p. 8).

Northpointe. (2015). Practitioner's guide to COMPAS core [https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf (Accessed: 2023-05-05)]. *Northpointe Inc.* (Cit. on p. 5).

Nyarko, J., Goel, S., & Sommers, R. (2021). Breaking taboos in fair machine learning: An experimental study. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11 (cit. on pp. 9, 14, 50, 54, 169, 171, 173, 174, 192).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453 (cit. on pp. 41, 43).

OECD. (2019). Recommendation of the Council on Artificial Intelligence [https://www.fsmb.org/siteassets/artificial-intelligence/pdfs/oecd-recommendation-on-ai-en.pdf (Accessed: 2023-06-19)]. *Organisation for Economic Co-operation and Development*. (Cit. on p. 28).

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13 (cit. on p. 41).

Ortega, A., Fierrez, J., Morales, A., Wang, Z., & Ribeiro, T. (2021). Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 78–87 (cit. on pp. 266, 267).

Padmanabhan, D., V, S., & Jose, J. (2020). On fairness and interpretability. *IJCAI 2020 AI for Social Good Workshop* (cit. on pp. 75, 267).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906 (cit. on p. 60).

Palan, S., & Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27 (cit. on pp. 110, 119, 142, 183).

Pan, W., Cui, S., Bian, J., Zhang, C., & Wang, F. (2021). Explaining algorithmic fairness through fairness-aware causal path decomposition. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1287–1297 (cit. on p. 267).

Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., & Pedreschi, D. (2021). FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management*, *58*(5), 102657 (cit. on pp. 266, 267).

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery*, *126*(2), 619–625 (cit. on p. 36).

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419 (cit. on p. 110).

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (cit. on pp. 66, 266).

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *29*(4), 35:1–35:33 (cit. on pp. 53, 172, 266).

Parafita, Á., & Vitrià, J. (2021). Deep causal graphs for causal inference, black-box explainability and fairness. *Artificial Intelligence Research and Development: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, 415–424 (cit. on pp. 266, 267).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253 (cit. on pp. 53, 172).

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022). Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22 (cit. on pp. 71, 75, 266, 267).

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021). Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15 (cit. on pp. 66, 266).

Pasquale, F. (2015). *The black box society*. Harvard University Press. (Cit. on p. 108).

Passi, S., & Vorvoreanu, M. (2022). Overreliance on AI: Literature review [https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf (Accessed: 2023-06-27)]. *Microsoft*. (Cit. on pp. 18, 23, 149–152, 171).

Pearson, K. (1911). On a correction to be made to the correlation ratio η. *Biometrika*, *8*(1/2), 254–256 (cit. on p. 90).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830 (cit. on pp. 89, 178).

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568 (cit. on pp. 84, 89, 96, 169, 171, 174, 192).

Peng, A., Nushi, B., Kiciman, E., Inkpen, K., & Kamar, E. (2022). Investigations of performance and bias in human-AI teamwork in hiring. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12089–12097 (cit. on p. 175).

Peters, F., Pumplun, L., & Buxmann, P. (2020). Opening the black box: Consumer's willingness to pay for transparency of intelligent systems. *ECIS 2020 Proceedings*, 90 (cit. on p. 136).

Petrović, A., Nikolić, M., Radovanović, S., Delibašić, B., & Jovanović, M. (2022). FAIR: Fair adversarial instance re-weighting. *Neurocomputing*, *476*, 14–37 (cit. on pp. 266, 267).

Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (cit. on pp. 54, 110, 113).

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., et al. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7), 736–745 (cit. on p. 174).

Plane, A. C., Redmiles, E. M., Mazurek, M. L., & Tschantz, M. C. (2017). Exploring user perceptions of discrimination in online targeted advertising. *Proceedings of the 26th USENIX Security Symposium*, 935–951 (cit. on pp. 54, 173).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, *30* (cit. on p. 50).

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52 (cit. on p. 172).

Pownall, C. (2019). AI, algorithmic and automation incident and controversy repository (AIAAIC) [https://www.aiaaic.org/home (Accessed: 2023-05-06)]. *AIAAIC*. (Cit. on p. 7).

Pradhan, R., Zhu, J., Glavic, B., & Salimi, B. (2022). Interpretable data-based explanations for fairness debugging. *Proceedings of the 2022 International Conference on Management of Data*, 247–261 (cit. on pp. 64, 69, 70, 267).

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (cit. on pp. 26, 170).

Prégent, A. (2022). Emotion recognition technology: Re-shaping human relationships. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 909–909 (cit. on p. 44).

Prolific. (2022). Why Prolific? [https://prolific.co/prolific-vs-mturk/ (Accessed: 2023-06-17)]. *Prolific*. (Cit. on p. 110).

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., & Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793 (cit. on pp. 34, 171, 192).

Quadrianto, N., Sharmanska, V., & Thomas, O. (2019). Discovering fair representations in the data domain. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8227–8236 (cit. on pp. 69, 266, 267).

Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S., & Pedreschi, D. (2020). Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, *54*(2), 425–437 (cit. on pp. 266, 267).

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on pp. 171, 172, 266).

Rädsch, T., Eckhardt, S., Leiser, F., Pandl, K. D., Thiebes, S., et al. (2021). What your radiologist might be missing: Using machine learning to identify mislabeled instances of X-ray images. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1294–1303 (cit. on p. 82).

Raff, E., Sylvester, J., & Mills, S. (2018). Fair forests: Regularized tree induction to minimize model bias. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 243–250 (cit. on pp. 68, 266, 267).

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14 (cit. on p. 52).

Raji, I. D., & Buolamwini, J. (2022). Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial AI products. *Communications of the ACM*, *66*(1), 101–108 (cit. on p. 29).

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 19–36 (cit. on pp. 66, 72, 266, 267).

Rawls, J. (1999). *A theory of justice: Revised edition*. Harvard University Press. (Cit. on pp. 38–40).

Räz, T. (2022). COMPAS: On a pathbreaking debate on algorithmic risk assessment. *Forensische Psychiatrie, Psychologie, Kriminologie*, *16*(4), 300–306 (cit. on pp. 4, 9).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788 (cit. on p. 31).

Reicin, E. (2021). AI can be a force for good in recruiting and hiring new employees [https://www.forbes.com/sites/forbesnonprofitcouncil/2021/11/16/ai-can-be-a-force-for-good-in-recruiting-and-hiring-new-employees/ (Accessed: 2023-05-06)]. *Forbes*. (Cit. on p. 3).

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112 (cit. on pp. 53, 172).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (cit. on pp. 7, 32, 170, 179, 209).

Riley, V. (2018). Operator reliance on automation: Theory and data. In *Automation and human performance: Theory and applications* (pp. 19–35). CRC Press. (Cit. on p. 172).

Robertson, J., Stinson, C., & Hu, T. (2022). A bio-inspired framework for machine bias interpretation. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 588–598 (cit. on p. 267).

Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705 (cit. on pp. 68, 266).

Rubenfeld, J. (1997). Affirmative action. *The Yale Law Journal*, *107*(2), 427–472 (cit. on p. 40).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215 (cit. on pp. 7, 10, 31, 33, 63, 73, 132, 266, 267).

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1), 1–55 (cit. on p. 6).

Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems*, *30* (cit. on pp. 66, 266, 267).

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach*. Pearson Education, Inc. (Cit. on p. 19).

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468 (cit. on p. 174).

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control. *Ethics and Information Technology*, *24*(4), 1–11 (cit. on p. 266).

Satariano, A. (2020). British grading debacle shows pitfalls of automating government [https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html (Accessed: 2023-06-17)]. *The New York Times*. (Cit. on pp. 3, 105, 138).

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., et al. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, *283*, 103238 (cit. on pp. 52, 53, 110).

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *ACM CHI 2022 Workshop on Trust and Reliance in AI-Assisted Tasks (TRAIT)* (cit. on pp. 151, 152, 171, 172, 177).

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626 (cit. on pp. 25, 34, 171, 209).

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422 (cit. on pp. 24, 53, 150–152, 155, 208).

Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. *ECIS 2022 Proceedings*, 51 (cit. on p. 172).

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. *Mensch und Computer 2021*, 325–329 (cit. on p. 109).

Schlicker, N., Langer, M., Ötting, S., Baum, K., König, C. J., et al. (2021). What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, *122*, 106837 (cit. on pp. 66, 111, 183, 266).

Schlicker, N., Uhde, A., Baum, K., Hirsch, M. C., & Langer, M. (2022). Calibrated trust as a result of accurate trustworthiness assessment – Introducing the trustworthiness assessment model. *PsyArXiv preprint 10.31234/osf.io/qhwvx* (cit. on pp. 209, 212).

Schöffer, J., De-Arteaga, M., & Kühl, N. (2022). On the relationship between explanations, fairness perceptions, and decisions. *ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI)* (cit. on p. 207).

Schöffer, J., Jakubik, J., Vössing, M., Kühl, N., & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. *HHAI 2023: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, 46–59 (cit. on pp. 171, 181).

Schöffer, J., & Kühl, N. (2021). Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 153–157 (cit. on pp. 66, 109, 145, 204, 209).

Schöffer, J., Kühl, N., & Machowski, Y. (2022). "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1616–1628 (cit. on p. 173).

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* (cit. on pp. 104, 140).

Schöffer, J., Ritchie, A., Naggita, K., Monachou, F., Finocchiaro, J., et al. (2023). Online platforms and the fair exposure problem under homophily. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(10), 11899–11908 (cit. on p. 205).

Schuetz, S. W., Steelman, Z. R., & Syler, R. A. (2022). It's not just about accuracy: An investigation of the human factors in users' reliance on anti-phishing tools. *Decision Support Systems*, *163*, 113846 (cit. on p. 151).

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*(3), 1085–1139 (cit. on pp. 56, 74, 75, 267).

Sen, A. (1979). Equality of what? [https://ophi.org.uk/wp-content/uploads/Sen-1979_Equality-of-What.pdf (Accessed: 2023-06-19)]. *The Tanner Lecture on Human Values*. (Cit. on p. 39).

Seymour, W. (2018). Detecting bias: Does an algorithm have to be transparent in order to be fair? *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems*, 2–8 (cit. on pp. 73, 266, 267).

Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 166–172 (cit. on pp. 63, 64, 68, 72, 75, 266, 267).

Shin, D. (2021a). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551 (cit. on p. 266).

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565 (cit. on p. 60).

Shin, D. (2021b). Why does explainability matter in news analytic systems? Proposing explainable analytic journalism. *Journalism Studies*, *22*(8), 1047–1065 (cit. on pp. 67, 266).

Shin, D., Lim, J. S., Ahmad, N., & Ibahrine, M. (2022). Understanding user sensemaking in fairness and transparency in algorithms: Algorithmic sensemaking in over-the-top platform. *AI & Society*, 1–14 (cit. on pp. 66, 266).

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284 (cit. on p. 172).

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2023). Enhancing fairness perception – Towards human-centred AI and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. *International Journal of Human-Computer Interaction*, *39*(7), 1455–1482 (cit. on pp. 64, 66, 72, 266).

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2022). Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*, *24*(1), 1–13 (cit. on pp. 54, 173, 266, 267).

Siegrist, M. (2008). Factors influencing public acceptance of innovative food technologies and products. *Trends in Food Science & Technology*, *19*(11), 603–608 (cit. on p. 113).

Siering, M. (2022). Explainability and fairness of RegTech for regulatory enforcement: Automated monitoring of consumer complaints. *Decision Support Systems*, *158*, 113782 (cit. on pp. 69, 267).

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–8 (cit. on p. 33).

Singh, A., & Joachims, T. (2017). Equality of opportunity in rankings. *Workshop on Prioritizing Online Content (WPOC) at NIPS*, *31* (cit. on p. 84).

Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2219–2228 (cit. on p. 84).

Skeem, J., Monahan, J., & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior*, *40*(5), 580–593 (cit. on p. 174).

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior*, *44*(1), 51–59 (cit. on p. 45).

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717 (cit. on pp. 134, 145).

Slack, D., Friedler, S. A., & Givental, E. (2020). Fairness warnings and Fair-MAML: Learning fairly with minimal data. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 200–209 (cit. on p. 267).

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, *34*, 62–75 (cit. on p. 267).

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186 (cit. on pp. 33, 74, 171, 206, 209, 267).

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, *1*(8), 330–331 (cit. on p. 51).

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285 (cit. on p. 113).

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1981). Perceived risk: Psychological factors and social implications. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *376*(1764), 17–34 (cit. on p. 113).

Smuha, N. (2018). A definition of AI: Main capabilities and scientific disciplines [https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (Accessed: 2023-06-27)]. *The European Commission's High-Level Expert Group on Artificial Intelligence*. (Cit. on p. 19).

Sokol, K., & Flach, P. (2019). Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. *SafeAI@AAAI* (cit. on pp. 68, 266).

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67 (cit. on p. 266).

Springer, A., & Whittaker, S. (2019). Making transparency clear. *Joint Proceedings of the ACM IUI 2019 Workshops* (cit. on pp. 72, 73, 170, 267).

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*, 2459–2468 (cit. on p. 136).

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 1–16 (cit. on pp. 7, 9, 34, 52–54, 173, 193, 207).

Steging, C., Renooij, S., & Verheij, B. (2021). Discovering the rationale of decisions: Towards a method for aligning learning and reasoning. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 235–239 (cit. on p. 266).

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for Kiva. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1241–1248 (cit. on pp. 68, 266).

Stumpf, S., Strappelli, L., Ahmed, S., Nakao, Y., Naseer, A., et al. (2021). Design methods for artificial intelligence fairness and transparency. *Joint Proceedings of the ACM IUI 2021 Workshops* (cit. on pp. 71, 267).

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9 (cit. on p. 42).

Szczygieł, K. (2022). In Poland, a law made loan algorithms transparent. Implementation is nonexistent [https://algorithmwatch.org/en/poland-credit-loan-transparency/ (Accessed: 2023-06-17)]. *AlgorithmWatch*. (Cit. on p. 107).

Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence*, *4*, 681915 (cit. on p. 266).

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273–1296 (cit. on p. 189).

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, *5*, 491–508 (cit. on pp. 24, 150, 261).

Thibaut, J. W., & Walker, L. (1975). *Procedural justice: A psychological analysis*. L. Erlbaum Associates. (Cit. on p. 112).

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*, 447–464 (cit. on p. 8).

Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 83–92 (cit. on pp. 68, 70, 266, 267).

Toussaint, P. A., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2022). Perceived fairness of direct-to-consumer genetic testing business models. *Electronic Markets*, *32*(3), 1621–1638 (cit. on p. 53).

Townson, S. (2020). AI can make bank loans more fair [https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair (Accessed: 2023-06-19)]. *Harvard Business Review*. (Cit. on pp. 3, 105, 108, 136).

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., et al. (2017). FairTest: Discovering unwarranted associations in data-driven applications. *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 401–416 (cit. on pp. 266, 267).

Treiss, A., Walk, J., & Kühl, N. (2021). An uncertainty-based human-in-the-loop system for industrial tool wear analysis. *ECML PKDD 2020: Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, 85–100 (cit. on p. 24).

Triberti, S., Durosini, I., & Pravettoni, G. (2020). A "third wheel" effect in health decision making involving artificial entities: A psychological perspective. *Frontiers in Public Health*, *8*, 117 (cit. on pp. 22, 106, 108).

Tsai, C.-H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and promoting diagnostic transparency and explainability in online symptom checkers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17 (cit. on p. 24).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460 (cit. on p. 19).

Uhde, A., Schlicker, N., Wallach, D. P., & Hassenzahl, M. (2020). Fairness and decision-making in collaborative shift scheduling systems. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on pp. 111, 112).

United States Congress. (2019). H.R. 5: Equality Act [https://www.congress.gov/bill/116th-congress/house-bill/5/text (Accessed: 2023-04-10)]. *United States Congress*. (Cit. on p. 83).

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. *Strategic HR Review*, *17*(5), 255–258 (cit. on p. 55).

van den Bos, K., Wilke, H. A., & Lind, E. A. (1998). When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology*, *75*(6), 1449–1458 (cit. on p. 112).

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404 (cit. on p. 172).

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., et al. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21 (cit. on pp. 173, 215, 266).

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of information presentation on fairness perceptions of machine learning predictors. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 266).

van Dongen, K., & van Maanen, P.-P. (2006). Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 225–229 (cit. on p. 25).

Vannur, L. S., Ganesan, B., Nagalapatti, L., Patel, H., & Tippeswamy, M. (2021). Data augmentation for fairness in personal knowledge base population. *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops*, 143–152 (cit. on pp. 266, 267).

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., et al. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38 (cit. on pp. 152, 155).

Velasquez, M., Andre, C., Shanks, T., & Meyer, M. J. (1990). Justice and fairness. *Issues in Ethics*, *3*(2), 1–3 (cit. on p. 37).

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39 (cit. on p. 108).

Vieira, C. P., & Digiampietri, L. A. (2022). Machine learning post-hoc interpretability: A systematic mapping study. *XVIII Brazilian Symposium on Information Systems*, 1–8 (cit. on p. 266).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272 (cit. on p. 91).

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2011). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. Springer Science & Business Media. (Cit. on p. 121).

von Zahn, M., Feuerriegel, S., & Kühl, N. (2022). The cost of fairness in AI: Evidence from e-commerce. *Business & Information Systems Engineering*, *64*, 335–348 (cit. on pp. 70, 99, 205).

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887 (cit. on pp. 33, 67, 73, 266, 267).

Wagner, B., & d'Avila Garcez, A. (2021). Neural-symbolic integration for fairness in AI. *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021)* (cit. on pp. 71, 267).

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, *11*(1), 104–122 (cit. on p. 23).

Waller, R. R., & Waller, R. L. (2022). Assembled bias: Beyond transparent algorithmic bias. *Minds and Machines*, *32*(3), 533–562 (cit. on pp. 70, 267).

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society*, *36*(2), 585–595 (cit. on pp. 74, 266, 267).

Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. Basic Books. (Cit. on p. 39).

Wang, C., Han, B., Patel, B., & Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, *39*, 519–581 (cit. on pp. 75, 266, 267).

Wang, G., Liu, X., Wang, Z., & Yang, X. (2020). Research on the influence of interpretability of artificial intelligence recommendation system on users' behavior intention. *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering*, 762–766 (cit. on p. 266).

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 54, 110, 113, 140).

Wang, S., & Gupta, M. (2020). Deontological ethics by monotonicity shape constraints. *International Conference on Artificial Intelligence and Statistics*, 2043–2054 (cit. on pp. 31, 85, 87, 132).

Wang, T., & Saar-Tsechansky, M. (2020). Augmented fairness: An interpretable model augmenting decision-makers' fairness. *arXiv preprint arXiv:2011.08398* (cit. on p. 267).

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257 (cit. on p. 44).

Wanner, J., Herm, L.-V., & Janiesch, C. (2020). How much is the black box? The value of explainability in machine learning models. *ECIS 2020 Proceedings*, 85 (cit. on p. 136).

Warner, R., & Sloan, R. H. (2021). Making artificial intelligence transparent: Fairness and the problem of proxy variables. *Criminal Justice Ethics*, *40*(1), 23–39 (cit. on pp. 69, 72, 73, 267).

Warren, R. C., Forrow, L., Hodge Sr, D. A., & Truog, R. D. (2020). Trustworthiness before trust – Covid-19 vaccine trials and the Black community. *New England Journal of Medicine*, *383*(22), e121 (cit. on p. 43).

Watson, D. S., & Floridi, L. (2021). The explanation game: A formal framework for interpretable machine learning. In *Ethics, governance, and policies in artificial intelligence* (pp. 185–219). Springer. (Cit. on pp. 67, 266, 267).

Wenar, L. (2021). John Rawls [https://plato.stanford.edu/entries/rawls/ (Accessed: 2023-06-30)]. *Stanford Encyclopedia of Philosophy*. (Cit. on p. 40).

Wilkinson, A., Roberts, J., & While, A. E. (2010). Construction of an instrument to measure student information and communication technology skills, experience and attitudes to e-learning. *Computers in Human Behavior*, *26*(6), 1369–1376 (cit. on pp. 116, 141).

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10 (cit. on pp. 59, 60).

Wohlin, C., Kalinowski, M., Felizardo, K. R., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology*, *147*, 106908 (cit. on p. 59).

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55 (cit. on p. 58).

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on p. 110).

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558 (cit. on p. 43).

Yang, K., & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1–6 (cit. on p. 84).

Yang, W., Lorch, L., Graule, M., Lakkaraju, H., & Doshi-Velez, F. (2020). Incorporating interpretable output constraints in Bayesian neural networks. *Advances in Neural Information Processing Systems*, *33*, 12721–12731 (cit. on p. 267).

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., et al. (2017). User trust dynamics: An investigation driven by differences in system performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317 (cit. on pp. 53, 172).

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731 (cit. on p. 55).

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, *20*(1), 2737–2778 (cit. on pp. 50, 83, 84, 173).

Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review*, *82*(3), 425–455 (cit. on p. 23).

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., et al. (2017). FA*IR: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, 1569–1578 (cit. on p. 84).

Zehlike, M., & Castillo, C. (2020). Reducing disparate exposure in ranking: A learning to rank approach. *Proceedings of The Web Conference 2020*, 2849–2855 (cit. on p. 84).

Zelaya, C. V. G. (2019). Towards explaining the effects of data preprocessing on machine learning. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2086–2090 (cit. on p. 267).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *International Conference on Machine Learning*, 325–333 (cit. on p. 49).

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*, 661–683 (cit. on p. 212).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340 (cit. on p. 50).

Zhang, H., Shahbazi, N., Chu, X., & Asudeh, A. (2021). FairRover: Explorative model building for fair and responsible machine learning. *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning*, 1–10 (cit. on pp. 266, 267).

Zhang, J., & Bareinboim, E. (2018). Fairness in decision-making – The causal explanation formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1) (cit. on pp. 63, 267).

Zhang, Y., & Ramesh, A. (2020). Learning fairness-aware relational structures. *ECAI 2020: 24th European Conference on Artificial Intelligence*, 2543–2550 (cit. on pp. 71, 267).

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305 (cit. on pp. 152, 171).

Zheng, H., Chen, Z., Du, T., Zhang, X., Cheng, Y., et al. (2022). NeuronFair: Interpretable white-box fairness testing through biased neuron identification. *Proceedings of the 44th International Conference on Software Engineering (ICSE 2022)*, 1519–1531 (cit. on p. 267).

Zhou, J., Chen, F., & Holzinger, A. (2020). Towards explainability for AI fairness. *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 375–386 (cit. on pp. 57, 64, 69, 267).

Zhu, Y.-Q., & Chen, H.-G. (2012). Service fairness and customer satisfaction in internet banking: Exploring the mediating effects of trust and customer value. *Internet Research*, *22*(4), 482–498 (cit. on p. 113).

Zippia. (2022a). Professor demographics and statistics in the US [https://www.zippia.com/professor-jobs/demographics/ (Accessed: 2023-06-19)]. *Zippia*. (Cit. on p. 175).

Zippia. (2022b). Teacher demographics and statistics in the US [https://www.zippia.com/teacher-jobs/demographics/ (Accessed: 2023-06-19)]. *Zippia*. (Cit. on p. 175).

Žliobaitė, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (cit. on p. 83).

Zucker, J., & d'Leeuwen, M. (2020). Arbiter: A domain-specific language for ethical machine learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 421–425 (cit. on pp. 72, 267).

Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. *Council of Europe, Directorate General of Democracy* (cit. on p. 42).

List of Figures

List of Tables

Appendix A

Tab. A.1.: Overview of canonical claims and prior work that has addressed them (1 of 2).

Claim	References
Transparency is a necessary or sufficient condition for fairness	Abdollahi and Nasraoui (2018), Alufaisan, Kantarcioglu, and Zhou (2021), Arrieta et al. (2020), Calegari, Ciatto, Denti, and Omicini (2020), Calegari, Ciatto, and Omicini (2020), Cath (2018), Colaner (2022), Ferreira and Monteiro (2020), Floridi et al. (2018), Galinkin (2022), Gill et al. (2020), Langer, Oster, et al. (2021), Leslie (2019), Mittelstadt et al. (2016), Nassih and Berrado (2020), Sartori and Theodorou (2022), Shulner-Tal et al. (2023), Sokol and Flach (2019), Steging et al. (2021), and Vieira and Digiampietri (2022)
Transparency increases stakeholders' fairness perceptions	Aïvodji et al. (2019), Angerschmid et al. (2022), Anik and Bunt (2021), Asher et al. (2022), Berscheid and Roewer-Despres (2019), Binns et al. (2018), Cath (2018), G. K. Y. Chan (2022), Cornacchia et al. (2021), Dodge et al. (2019), Floridi et al. (2018), Galhotra et al. (2021, 2017), Gill et al. (2020), Gilpin et al. (2018), Gupta et al. (2019), Hacker and Passoth (2020), Hind et al. (2019), John-Mathews (2022), Karimi et al. (2022), Kizilcec (2016), Le Merrer and Trédan (2020), M. K. Lee et al. (2019), Loi et al. (2021), Nakao et al. (2022), Papenmeier et al. (2019, 2022), Park et al. (2022, 2021), Rader et al. (2018), Ras et al. (2018), Rudin (2019), C. Russell et al. (2017), Schlicker et al. (2021), Shin (2021a, 2021b), Shin et al. (2022), Shulner-Tal et al. (2023, 2022), Sokol and Flach (2020), van Berkel et al. (2021), Wachter et al. (2018), Walmsley (2021), G. Wang et al. (2020), and Watson and Floridi (2021)
Transparency enables stakeholders to assess fairness	Abdollahi and Nasraoui (2018), A. Aggarwal et al. (2019), Ahn and Lin (2020), Aïvodji et al. (2021), Alikhademi et al. (2021), Alves, Bhargava, et al. (2021), Alves et al. (2020), Anders et al. (2022), Balkir et al. (2022), Begley et al. (2020), Berscheid and Roewer-Despres (2019), Black et al. (2020), Borrellas and Unceta (2021), Castelnovo et al. (2021), Cesaro and Gagliardi Cozman (2019), Chung et al. (2019), Dash et al. (2022), Datta et al. (2016), de Greeff et al. (2021), Dimanov et al. (2020), Doshi-Velez and Kim (2017), Du et al. (2021), Fan et al. (2022), Franco et al. (2022), Franke (2022), Gill et al. (2020), Gilpin et al. (2018), Grabowicz et al. (2022), Hacker and Passoth (2020), Hall and Gill (2017), Hardt et al. (2021), Hind et al. (2019), Hohman et al. (2019), Ignatiev et al. (2020), Jain et al. (2020), Kroll et al. (2017), Langer, Oster, et al. (2021), Lipton (2018), Manerba and Guidotti (2021), Meng et al. (2022), Michael (2019), Miron et al. (2021), Miroshnikov et al. (2022), Ortega et al. (2021), Panigutti et al. (2021), Papenmeier et al. (2019), Parafita and Vitrià (2021), Petrović et al. (2022), Quadrianto et al. (2019), Qureshi et al. (2020), Raff et al. (2018), Ras et al. (2018), Rosenfeld and Richardson (2019), Rudin (2019), Seymour (2018), Sharma et al. (2020), Sokol and Flach (2019), Stevens et al. (2020), Szepannek and Lübke (2021), Tolan et al. (2019), Tramer et al. (2017), van Berkel et al. (2019), Vannur et al. (2021), Vieira and Digiampietri (2022), C. Wang et al. (2022), and H. Zhang et al. (2021)

Tab. A.2.: Overview of canonical claims and prior work that has addressed them (2 of 2).

Claim	References
Transparency enables stakeholders to understand sources of unfairness	Abdollahi and Nasraoui (2018), A. Aggarwal et al. (2019), Ahn and Lin (2020), Alikhademi et al. (2021), Anders et al. (2020), Arrieta et al. (2020), Begley et al. (2020), Black et al. (2020), Chakraborty et al. (2020), Datta et al. (2016), de Greeff et al. (2021), Du et al. (2021), Galhotra et al. (2021), Ge et al. (2022), A. Ghosh et al. (2022), B. Ghosh et al. (2023), Gill et al. (2020), Grabowicz et al. (2022), Leslie (2019), Manerba and Guidotti (2021), Miron et al. (2021), Miroshnikov et al. (2022), Mutlu et al. (2022), Ortega et al. (2021), Pan et al. (2021), Pradhan et al. (2022), Quadrianto et al. (2019), Qureshi et al. (2020), Raff et al. (2018), Robertson et al. (2022), Siering (2022), Slack, Hilgard, et al. (2020), Tolan et al. (2019), Tramer et al. (2017), Vannur et al. (2021), Warner and Sloan (2021), Zelaya (2019), H. Zhang et al. (2021), J. Zhang and Bareinboim (2018), Zheng et al. (2022), and Zhou et al. (2020)
Transparency enables stakeholders to mitigate unfairness	Aghaei et al. (2019), Ahn and Lin (2020), Aïvodji et al. (2021), Alves, Amblard, et al. (2021), Alves, Bhargava, et al. (2021), Alves et al. (2020), Bhargava et al. (2020), Bhatt et al. (2021), Chakraborty et al. (2020), G. K. Y. Chan (2022), Dash et al. (2022), de Greeff et al. (2021), Fan et al. (2022), Floridi et al. (2020), Franke (2022), Ge et al. (2022), B. Ghosh et al. (2023), Grabowicz et al. (2022), Gupta et al. (2019), Hall and Gill (2017), Hall et al. (2019), Hamon et al. (2022), Hickey et al. (2021), Hind et al. (2019), Jain et al. (2020), Karimi et al. (2022), Kung and Yu (2020), Langer, Oster, et al. (2021), Lawless et al. (2021), Linardatos et al. (2020), Maclure (2021), Meng et al. (2022), Miroshnikov et al. (2022), Mutlu et al. (2022), Pan et al. (2021), Panigutti et al. (2021), Parafita and Vitrià (2021), Park et al. (2022), Pradhan et al. (2022), C. Russell et al. (2017), Siering (2022), Slack, Friedler, and Givental (2020), Stumpf et al. (2021), Tolan et al. (2019), Wagner and d'Avila Garcez (2021), Waller and Waller (2022), T. Wang and Saar-Tsechansky (2020), W. Yang et al. (2020), H. Zhang et al. (2021), J. Zhang and Bareinboim (2018), Y. Zhang and Ramesh (2020), Zheng et al. (2022), and Zucker and d'Leeuwen (2020)
Transparency enables stakeholders to certify fairness	Adadi and Berrada (2018), G. K. Y. Chan (2022), Cornacchia et al. (2021), de Fine Licht and de Fine Licht (2020), Dodge et al. (2019), Floridi et al. (2018), Grimmelikhuijsen (2023), Gryz and Shahbazi (2020), Hacker and Passoth (2020), Hall et al. (2019), Hamon et al. (2022), Hickey et al. (2021), Hind et al. (2019), John-Mathews (2022), Kehrenberg et al. (2020), Kroll et al. (2017), Le Merrer and Trédan (2020), Lepri et al. (2018), Leslie (2019), Loi et al. (2021), Maclure (2021), Padmanabhan et al. (2020), Ras et al. (2018), Seymour (2018), Sharma et al. (2020), Shulner-Tal et al. (2022), Springer and Whittaker (2019), Wachter et al. (2018), Walmsley (2021), Warner and Sloan (2021), Watson and Floridi (2021), and Zucker and d'Leeuwen (2020)
Transparency interventions are prone to misinterpretation and manipulation	Aïvodji et al. (2021, 2019), Anders et al. (2020), Balagopalan et al. (2022), Begley et al. (2020), Chakraborty et al. (2020), Dai et al. (2022), Dimanov et al. (2020), Galinkin (2022), Gill et al. (2020), Gilpin et al. (2018), Herman (2017), Kasirzadeh and Smart (2021), Lipton (2018), Rudin (2019), Selbst and Barocas (2018), Slack, Hilgard, et al. (2020), Slack et al. (2021), and Walmsley (2021)
Transparency and fairness are conflicting goals	Adomavicius and Yang (2022), Aghaei et al. (2019), Aïvodji et al. (2021), Begley et al. (2020), Borrellas and Unceta (2021), Du et al. (2021), Floridi et al. (2020), Gryz and Shahbazi (2020), He et al. (2020), Hickey et al. (2021), Jabbari et al. (2020), Kleinberg and Mullainathan (2019), Kung and Yu (2020), Lawless et al. (2021), Padmanabhan et al. (2020), Petrović et al. (2022), Rudin (2019), Selbst and Barocas (2018), Walmsley (2021), C. Wang et al. (2022), T. Wang and Saar-Tsechansky (2020), and Y. Zhang and Ramesh (2020)