Mapping the Potential of Explainable Artificial Intelligence (XAI) for Fairness Along the AI Lifecycle

LUCA DECK, University of Bayreuth & Fraunhofer FIT, Germany
ASTRID SCHOMÄCKER, University of Bayreuth, Germany
TIMO SPEITH, University of Bayreuth, Germany
JAKOB SCHÖFFER, University of Texas at Austin, USA
LENA KÄSTNER, University of Bayreuth, Germany
NIKLAS KÜHL, University of Bayreuth & Fraunhofer FIT, Germany

The widespread use of artificial intelligence (AI) systems across various domains is increasingly highlighting issues related to algorithmic fairness, especially in high-stakes scenarios. Thus, critical considerations of how fairness in AI systems might be improved, and what measures are available to aid this process, are overdue. Many researchers and policymakers see explainable AI (XAI) as a promising way to increase fairness in AI systems. However, there is a wide variety of XAI methods and fairness conceptions expressing different desiderata, and the precise connections between XAI and fairness remain largely nebulous. Besides, different measures to increase algorithmic fairness might be applicable at different points throughout an AI system's lifecycle. Yet, there currently is no coherent mapping of fairness desiderata along the AI lifecycle. In this paper, we set out to bridge both these gaps: We distill eight fairness desiderata, map them along the AI lifecycle, and discuss how XAI could help address each of them. We hope to provide orientation for practical applications and to inspire XAI research specifically focused on these fairness desiderata.

1 INTRODUCTION

The emergence and widespread use of artificial intelligence (AI) systems across various sectors and domains is increasingly shifting attention from considerations of mere performance to considerations about algorithmic fairness. This is particularly relevant for systems employed in high-stake scenarios and especially pressing in contexts prone to harmful societal biases [14, 103]. Thus, critical considerations of how fairness in AI systems might be improved, and what measures are available to aid this process, are overdue. In the literature, there is a common recognition that fairness in AI systems demands various perspectives and measures (e.g., [6, 35, 128]). In this paper, we set out to integrate two strategies that have been suggested: First, a growing community of researchers proposes explainable AI (XAI) as a versatile and powerful tool to combat unfairness [36]. Second, others have focused on the AI lifecycle and tried to determine where fairness issues originate (e.g., [3, 48]). Their hope is that, once identified, issues with fairness can be mitigated by taking appropriate steps in the relevant phase within the lifecycle.

While both these strategies seem intuitively promising, neither currently presents a satisfactory and comprehensive picture. A primary reason why neither approach currently fulfills their potential is, we take it, that there are various types and kinds of fairness discussed in the literature, yielding different *fairness desiderata*. Differentiating between these desiderata is crucial to gain an overall picture of which measures are most promising to address fairness in which contexts. Without such differentiation, the utility of XAI is not as straightforward as commonly claimed. Accordingly, there is a need for more clarity about how, exactly, XAI contributes to which fairness desideratum [36]. Existing discussions on this matter suffer, by and large, from both a too narrow conception of XAI and a mostly under-specified

Authors' addresses: Luca Deck, luca.deck@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Germany; Astrid Schomäcker, astrid.schomaecker@uni-bayreuth.de, University of Bayreuth, Germany; Timo Speith, timo.speith@uni-bayreuth.de, University of Bayreuth, Germany; Jakob Schöffer, schoeffer@utexas.edu, University of Texas at Austin, USA; Lena Kästner, lena.kaestner@uni-bayreuth.de, University of Bayreuth, Germany; Niklas Kühl, kuehl@uni-bayreuth.de, University of Bayreuth & Fraunhofer FIT, Germany.

notion of fairness. Similarly, existing attempts to map measures for achieving fairness onto the AI lifecycle remain crucially incomplete; for they have limited their attention to only a subset of the relevant fairness desiderata. To overcome these limitations, we distill eight fairness desiderata, map them along the AI lifecycle, and discuss how XAI can help address each of them.

For the purposes of this paper, we will focus specifically on how XAI can be utilized to improve AI systems' fairness throughout their lifecycle. With respect to this question, we aim to develop a holistic account based on the fairness desiderata we distill. While our proposal is not intended as a guideline, we believe it will stimulate scientific discussions about and practical applications of fairness measures; and as such will be a valuable starting point to explore fairness opportunities for researchers, developers, and regulators alike.

We begin by introducing preliminaries and diagnosing the problem in Section 2. Importantly, we do not limit our understanding of algorithmic fairness to computational perspectives. In Section 3 we propose eight fairness desiderata: fairness understanding, data fairness, formal fairness, perceived fairness, fairness with human oversight, empowering fairness, long-term fairness, and informational fairness (contribution #1). We offer a mapping of our fairness desiderata along the different stages within the AI lifecycle and suggest that each fairness desideratum affords a different entry point for taking measures to improve fairness. This closes the first gap in the current literature as it highlights where in the lifecycle different fairness desiderata become especially relevant (contribution #2). Utilizing this mapping, we discuss how XAI can be leveraged to address the different fairness desiderata at the respective points throughout the AI lifecycle. This closes the second gap in the current literature as it allows to systematically examine the potential of XAI to address algorithmic fairness in different circumstances (contribution #3). To illustrate the utility of our approach, we discuss its application to the COMPAS case (see [7]) in Section 4. Before closing, we point to some avenues for future research in Section 5.

2 BACKGROUND

In this article, we contribute to closing two research gaps: 1) considerations of fairness along the AI lifecycle are incomplete, and 2) it is unclear how exactly XAI can help in fostering fairness. Before closing these gaps, we retrace contributions and debates of prior works.

2.1 Algorithmic Fairness

The debate on algorithmic fairness draws inspiration from various disciplines. Several scandals surrounding AI systems that disadvantage marginalized groups [7, 34, 51] have shattered early hopes put into the "neutrality" of AI [14]. In reaction, scholars from computer science, philosophy, social science, law, and psychology have been engaging in debates on what it means for algorithmic decision-making to be fair (see [99] for interdisciplinary perspectives).

The technical debate on algorithmic fairness primarily focuses on formal measures of fairness [13, 23, 132]. After it has proven unhelpful to remove information about membership in marginalized groups from training data ("fairness through unawareness" [42]), most approaches from the field of computer science rely on comparing the outcomes for people from different groups. An important finding is that the different formal measures for fairness can, in most cases, not be fulfilled simultaneously [37, 72]. Thus, (formally) fair AI systems require crucial design choices about which formal measures to apply in which context.

To aid such decisions, philosophers have connected formal measures to different philosophical theories [17, 71, 85]. In general, "fairness" is a normative concept—and debates about fairness can be understood as a way to discuss what is

morally right or wrong in a given case (see [99]). More specifically, fairness can be understood as an issue of justice, non-discrimination, or equality [17], which connects it to numerous strands of philosophical discussion.

Beyond that, social scientists have pointed out shortcomings of existing formal methods, e.g., problems in detecting issues of intersectionality [33] and their reliance on constructed categories like race and gender [64]. Legal scholars are primarily interested in how legislation on discrimination can be applied to algorithmic decision-making and whether additional regulation is required [14, 135]. Psychological research is generally interested in whether algorithmic decision-making is perceived as fair [82, 128]. For example, Colquitt and Rodell [31] distinguish four different psychological dimensions of fairness: Whether the outcome is fair (viz., distributive justice), whether the process that leads to the outcome is fair (viz., procedural justice), whether the information about the decision is communicated truthfully and thoroughly (viz., informational justice), and whether individuals are treated respectfully (viz., interpersonal fairness).

Overall, algorithmic fairness can be addressed from several different angles—some of which are complimentary, some of which are contradictory. This stresses the need to be clear about the specific desideratum behind any attempt to improve fairness.

2.2 Considerations of Fairness Along the AI Lifecycle Remain Incomplete

A common strategy to determine potential entry points to address fairness issues involves examining each step of the AI lifecycle. There have been several proposals of prototypical AI lifecycles (e.g., [47, 74, 110, 137, 139]), each with slightly different stages or adapted to different application areas. The AI lifecycle used in this article is a combination of the proposals by Quemy [110] and Wang et al. [137] involving the following stages: 1) problem formulation, 2) data collection, 3) data analysis, 4) feature selection, 5) model construction, 6) model evaluation, 7) deployment, and 8) inference and usage (see Figure 1).

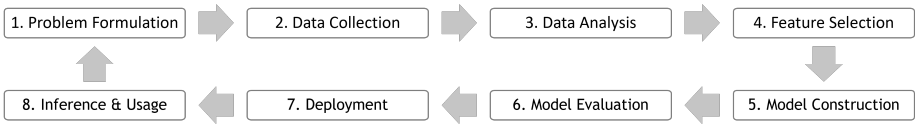

Fig. 1. The AI Lifecycle we use, combined from Quemy [110] and Wang et al. [137].

First, *problem formulation* involves abstraction and formulation of the problem, which is to be solved using AI. Afterward, *data collection* aims to establish a representative data set suitable to develop an AI system able to solve the defined problem. *Data analysis* includes descriptive statistics to understand the characteristics of the data at hand and data pre-processing to prepare the data for further operations. Subsequently, in the *feature selection* stage, features are excluded, transformed, or aggregated for effective and efficient handling in the training process. *Model construction*, then, includes selection of the training algorithm and the training itself. Iteratively, performance objectives are tested during *model evaluation* and optimized through adjustments to the algorithm, e.g., in the form of parameter tuning. *Deployment* refers to the integration of the AI system into a productive environment once the model achieves sufficient performance. Finally, *inference & usage* describes concrete output generated by the AI system and its impact on the surroundings, such as the business context, society, and environment. Note that these stages are often not entered sequentially but leave room for iterations and loops between stages.

Much research has focused on what types of bias can emerge in or affect different steps of the AI lifecycle (e.g., [3, 35, 86, 125]). For example, Singh et al. [125] take the CRISP-DM model [139] and review the types of bias that can

occur at different stages of this model. However, these considerations remain incomplete, as most of these works have a strong technical focus and deal only with formal notions of fairness, e.g., by providing guidance for choosing fairness metrics or debating the role of potential fairness-utility tradeoffs. Against this background, it is important that we develop a view that also maps other conceptions of fairness into the lifecycle.

2.3 Considerations of XAI for Fairness Remain Incomplete

To amend the lack of understanding of AI-based systems, their reasoning processes, and their outputs, the research field of XAI emerged in recent years [15, 26, 73, 79, 105]. Generally, the goal of XAI is to provide a stakeholder with information about some aspect of an AI system to facilitate their understanding of this aspect [26]. Understanding (some aspect of) a system is often just an intermediary step to other goals, such as fairness or appropriate reliance [78, 120]. For example, by understanding how a particular system's output came to be, prior work has argued that a person should be enabled to assess whether this output was based on valid criteria or not [56, 105]. If an unfavorable decision was based on (a proxy for) the skin color of a person, this person should supposedly be able to recognize that they were treated unfairly [22, 77].

However, the goal of XAI is often narrowly construed as the development of methods to create interpretable surrogate models of black box models [41, 115]. Such a narrow view of XAI artificially restricts the space of options that can be chosen to facilitate stakeholders' understanding of AI systems. Take, e.g., model cards for model reporting [98] or datasheets for data sets [52]. Although these approaches provide important information about certain aspects of an AI model (e.g., which training data was used) and thus improve stakeholders' understanding of these aspects, they would rarely be counted as XAI. Further, XAI methods alone do not provide other kinds of highly relevant information, such as the normative motivations guiding the development of an AI model [93], the analysis of the social context in which it is deployed [58], or the descriptive outcome statistics of the deployed model [29].

In a similar vein, the high hopes placed in XAI to mitigate issues of fairness often remain vague or unfulfilled [36]. Prior works have criticized the explanatory value [67, 115], susceptibility to manipulations [5, 75], and unsatisfactory interpretations [11, 36] of XAI methods. For example, popular feature-based explanations like LIME [114] or SHAP [94] highlighting the use of sensitive features (e.g., gender and race) provides little information about fairness because these features are often correlated with proxy variables and embedded in use case-specific normative contexts [42, 91, 104]. Instead, broadening the conception of XAI enables a more holistic and meaningful mapping of how information about an AI system can contribute to various fairness desiderata.

Against this background, we propose to distinguish the *narrow* conception of XAI, solely pertaining to the inner workings of an AI system, from a *broad* conception of XAI that provides information beyond these inner workings and includes explanations of the broader socio-technical system [38]. Specifically, we suggest that the broad conception of XAI includes all types of information that increases stakeholders' understanding of (aspects of) an AI system. If not indicated otherwise, we will simply refer to XAI and mean the broad conception.

3 HOW XAI CAN BE LEVERAGED FOR FAIRNESS ALONG THE AI LIFECYCLE

Grounded in a broad interdisciplinary literature review (summarized in Table 1), we distill eight categories of what previous work on fairness has called for. We call these categories *fairness desiderata*, and each desideratum can be instantiated with specific *fairness objectives*. For the purposes of this paper, we focus specifically on fairness desiderata connected to XAI. While there are various stakeholders involved in utilizing XAI for fairness, an extensive discussion of

their various roles is beyond the scope of this paper. When we mention stakeholders, we rely on the taxonomy provided by Langer et al. [79].

In what follows, we introduce our eight fairness desiderata, describe the underlying core ideas for each, discuss how they relate to similar concepts, map them onto the AI lifecycle, and elaborate on how we think XAI could contribute to their satisfaction (see Figure 2). Note that our proposal does not necessarily present a comprehensive list (see Section 5).

Table 1. Fairness desiderata and their related concepts in existing interdisciplinary literature.

Fairness desiderata	Related concepts
Fairness understanding Gaining higher-level insights on fairness and the socio-technical challenges surrounding the development and deployment of an AI application to specify concrete fairness objectives.	Understanding bias [103] Interdisciplinary fairness conceptualization [99] Lessons from political philosophy [17] Socio-technical perspective [117]
Data fairness Identifying and addressing flaws in the data set that might be unfair themselves or potentially lead to downstream violations of fairness objectives.	Sampling bias [35, 96] Data errors and bias [109] Data-centric factors in algorithmic fairness [88]
Formal fairness Identifying and addressing model properties leading to violations of formal fairness objectives.	Algorithmic bias [35] Formal fairness definitions [32] Fairness metrics [23] Statistical fairness criteria [13] Disparate impact & disparate treatment [14]
Perceived fairness Providing affected parties with explanations and justifications to improve or "calibrate" fairness perceptions.	Fairness perceptions [122, 128] Perceptions of justice [19, 30] Fairness judgment [39] Society-in-the loop [111]
Fairness with human oversight Supporting human decision-makers interacting with an AI system to effectively align human discretion with fairness objectives.	Human-ML augmentation for fairness [131] Appropriate reliance [118] Human-in-the-loop [39] Domain expertise [25]
Empowering fairness Providing affected parties with practical information to foster contestability and recourse.	Self-informed advocacy [133] Procedural fairness [45] Counterfactual explanations [134] Fair and adequate explanations [10] The explanation game [138]
Long-term fairness Monitoring and analyzing the socio-technical long-term impacts of algorithmic decision-making to adjust unfair repercussions over time.	Long-term effects of algorithmic fairness [35] Fairness drift [53] Fairness through time [24] Fairness monitoring [65, 106]
Informational fairness Providing truthful, understandable, and relevant information about all fairness desiderata across the AI lifecycle.	Informational fairness [30] Design publicity [93] Model cards [98] Outward transparency [136]

3.1 Fairness Understanding

> *"Accounting for bias not only requires an understanding of the different sources, that is, data, knowledge bases, and algorithms, but more importantly, it demands the interpretation and description of the meaning, potential side effects, provenance, and context of bias."* [103, p. 8]

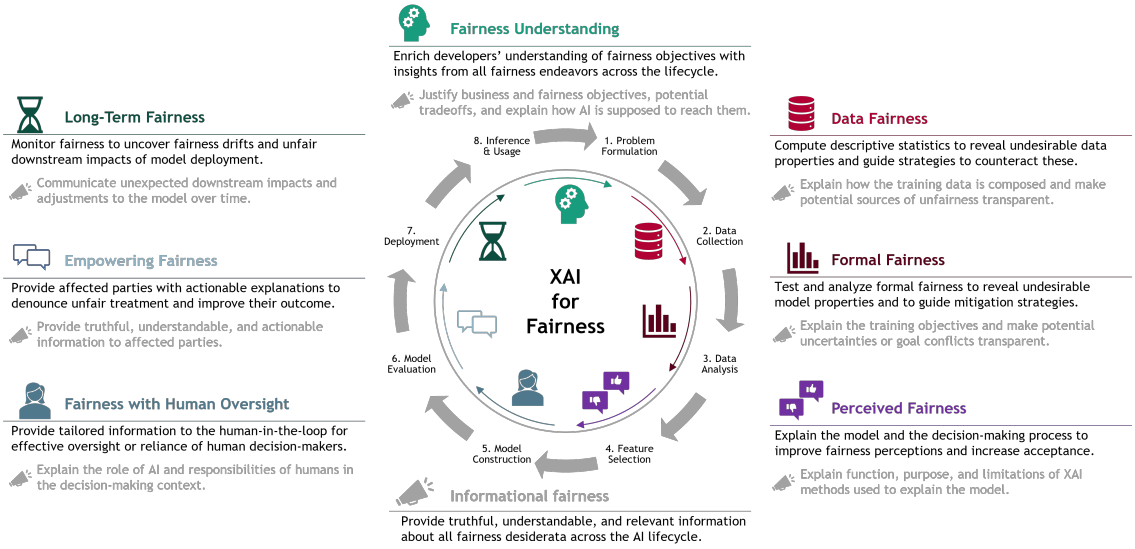

Fig. 2. Fairness desiderata along the AI lifecycle depicting how XAI may directly contribute to these desiderata and how it may contribute to informational fairness across all desiderata (symbolized by speakerphones).

Before mitigating algorithmic unfairness, developers should reflect upon the multidimensional and conflicting nature of fairness, define concrete fairness objectives, and understand how these might be achieved. In light of this, *fairness understanding* relates to the specification of concrete fairness objectives and knowledge about the socio-technical challenges surrounding the development and deployment of an AI application.

Among others, this may include a sound understanding of fairness schools of thought [99], existing social inequalities [58], stakeholder-centered fairness requirements [79], or relevant legal frameworks [63]. As the following fairness desiderata will show, there is no one-size-fits-all way in which a system can be fair, both across (e.g., a system can be perceived as fair despite not being formally fair) as well as within fairness desiderata (e.g., formal fairness criteria pose inherent tradeoffs [49]).

Stage of the Lifecycle. In the AI lifecycle, we map fairness understanding to the initial problem formulation step because all subsequent steps are guided by the fairness objectives discussed and defined in the early stages. We note, however, that fairness understanding is developed iteratively, based on trials and errors, interdisciplinary exchange, and stakeholder feedback from various development stages—similarly to how a traditional business problem can be better understood and adjusted over the entire course of an AI project.

The Role of XAI. XAI can contribute to a better understanding of fairness across all fairness desiderata which, e.g., may lead to a re-evaluation of earlier fairness objectives. Data-centric explanations [8] might reveal pre-existing group disparities (e.g., in the form of differing base rates such as the gender pay gap), which are crucial factors in determining the fairness objectives and how to achieve them. Further, simulating and testing various formal fairness metrics promotes an understanding of conflicting objectives that may lead to adjustments or re-weightings of fairness objectives [35]. Similarly, XAI may change stakeholders' view of proactively using sensitive features in a specific decision-making context [104]. With regard to "substantive" fairness [58, 71], XAI may also be applied to normatively gauge the legitimacy of certain features depending on the societal context.

3.2 Data Fairness

> *"Explanations [...] are crucial for helping system developers and ML practitioners to debug ML algorithms by identifying data errors and bias in training data, such as measurement errors and misclassifications, data imbalance, missing data and selection bias, covariate shift, technical biases introduced during data preparation, and poisonous data points injected through adversarial attacks."* [109, p. 248]

Undesirable model behavior often stems from flawed data that contains misrepresentations of the world (e.g., erroneous, mislabeled, or imbalanced data) or accurate representations that are societally undesirable (e.g., historical inequalities). The goal of *data fairness* is to identify and address such flaws in the data set used for the training of an AI model.

Prior studies have pointed out data issues as drivers of unfairness [12, 88, 96] or highlighted data as a starting point for unfairness mitigation [21, 89, 116]. Mehrabi et al. [96] provide a comprehensive list of data biases that may introduce unfairness early on in AI development. Awareness and understanding of these biases are crucial to derive strategies to handle them.

Stage of the Lifecycle. Data fairness relates to data collection (e.g., in the form of labeling errors, sampling bias, imbalances, etc.) and data analysis, which aims to identify and potentially mitigate unfair data characteristics early on. However, data fairness also reaches into feature selection which is usually part of an iterative loop together with model construction. For example, features might be dropped if they are not justifiably task-relevant. Importantly, in the context of sensitive attributes, developers should always be aware of the flaws of the idea of "fairness through unawareness" [42, 104].

The Role of XAI. As data is a primary source of unfairness, XAI is suited to identify potential disparities, imbalances, or abnormalities manifested in the available data early on. Descriptive statistics are a natural first step to explore pre-existing disparities [88]. Anik and Bunt [8] and Mitchell et al. [98] provide two examples of how to present simple descriptions and visualizations about the collection, feature distributions, and patterns of a dataset. XAI techniques have further been claimed to reveal instances and features in the data that have undesirable effects on the model output [20, 46, 109] which, however, are often subject to questionable causal and normative assumptions [18, 28]. These approaches indicate a strong connection between data fairness and *formal fairness*.

3.3 Formal Fairness

> *"To counteract biases, it is, therefore, crucial to enable their detection. Explainability approaches may aid in this regard by providing means to track down factors that may have contributed to unfair and unethical decision-making processes and either to eliminate such factors, to mitigate them, or at least to be aware of them."* [79, p. 6]

The most common fairness desideratum is concerned with formal model properties. By *formal fairness*, we refer to the vast array of formal criteria that have been proposed as mathematical and statistical measures of fairness [13, 23].

Consistent with Verma and Rubin [132], this includes all fairness definitions based on statistical measures (e.g., demographic parity), similarity-measures (e.g., fairness through awareness), or based on causal reasoning (e.g., counterfactual fairness). Formal fairness notions are often distinguished into group and individual fairness. Group fairness criteria typically require a form of parity between demographic groups, e.g., along sensitive attributes like gender or race [29]. Individual fairness criteria typically demand to treat similar people alike [42].

Stage of the Lifecycle. Formal fairness is particularly relevant for the iterative loop of model construction and model evaluation. As soon as the first model prototype is ready, fairness metrics can be evaluated. Based on the evaluation, unfairness mitigation techniques can be implemented and validated iteratively [27].

The Role of XAI. Regarding formal fairness, XAI could be of exploratory value offering a plethora of tools to explore formal fairness from multiple angles. Again, descriptive statistics present a natural first step to identify potential disparities, imbalances, or abnormalities outputted by the model [4, 29]. This is especially useful if formal fairness objectives are already specified, e.g., when testing whether formal group fairness metrics are sufficiently satisfied [27]. XAI might also point towards specific features driving the violation of group fairness metrics [54] or shed light on more subtle forms of formal fairness such as fairness of recourse [61]. XAI has further been claimed to elucidate the complex interplay between "task-relevant" and correlated "protected" features (e.g., [57]) which, again, relies on causal and normative assumptions [18, 28]. Technical possibilities of XAI to explore formal fairness are manifold but require utmost caution when applied for unfairness mitigation in specific application contexts and should not be misinterpreted as fairness "proofs" or "guarantees" [36]. Particularly, when interpreting the legitimacy of using sensitive information, it is paramount to account for the "fairness through unawareness" fallacy [42, 91] and for differential subgroup validity [60, 69].

3.4 Perceived Fairness

> *"We argue that only fair systems that are also perceived to be fair by their users should be accepted and employed in practice."* [123, p. 5]

Another common fairness desideratum refers to positive perceptions of stakeholders, which is often related to trust and acceptance [107, 123]. In this sense, *perceived fairness* captures how stakeholders, particularly those affected by a decision, perceive the fairness of the AI system.

Measures of perceived fairness can be derived from the justice constructs of Colquitt [30], which decompose fairness perceptions into a procedural, distributive, and informational dimension (see also [19, 39, 122]). Accounting for fairness perceptions promotes a valuable means to design AI systems based on human needs and ideals while giving voice to societal values which Rahwan [111] coined as "society-in-the-loop". Notably, the desirable aspect of perceived fairness does not necessarily or solely lie in positive fairness perceptions but in "appropriate" fairness perceptions [121].

Stage of the Lifecycle. Although fairness perceptions can be relevant at any stage of the lifecycle (e.g., to evaluate data fairness [8]), we map fairness perceptions to the model evaluation, deployment, and inference & usage stage. Most prior works have measured fairness perceptions after a (mock-up) model has been developed to ask stakeholders about certain outcomes or the AI system itself [128]. Beyond that, there are approaches trying to embed stakeholders' values throughout conception and development of AI systems (e.g., [84, 102, 130, 140]).

The role of XAI. Explanations may serve stakeholders as a cue to confirm whether their ideas of fairness are implemented in a system. This idea has also been commonly expressed in prior literature [107, 113, 123, 124]. For example, Lee et al. [83] indicate that explanations can decrease fairness perceptions when they reveal information that stands in conflict with peoples' fairness beliefs. However, the effect of XAI on fairness perceptions is highly context-dependent and moderated by human factors like political ideology and self-interest [129]. This indicates that stakeholders require tailored information addressing their case-specific concerns. Affected parties, e.g., seem to appreciate information beyond a model's inner workings such as system context, usage, and data [119]. We note that optimizing XAI to stimulate positive fairness perceptions isolated from complementary desiderata (e.g., trustworthiness or formal fairness) can lead to undesirable effects such as placebic explanations [43], fairwashing [5] or deception [70, 80].

3.5 Fairness With Human Oversight

> *"Research suggests that neither humans nor ML models are likely to achieve fairness working alone. Instead, human–ML augmentation, where humans and technology work together to perform organizational tasks jointly, is the most promising path to achieving fairness."* [131, p. 2]

Beyond the (formal) fairness of a model itself, fairness can also refer to the decision-making process the model is embedded in. *Fairness with human oversight* aims at installing and supporting a human decision-maker to realize case and context-sensitive fairness objectives through human oversight or human discretion.

The human-AI setting can take various forms but usually involves overseeing and overruling unfair outputs or fostering effective reliance behavior [131]. Although legal and ethical guidelines often demand human oversight [44, 45], the effect of human oversight on fairness is not necessarily beneficial and not well understood yet [76, 120, 131, 136]. Because full automation is currently not permitted in many contexts, this desideratum does not necessarily capture fairness *through* human oversight, but potentially also fairness *despite* human oversight.

Stage of the Lifecycle. Fairness with human oversight becomes relevant after the model has been deployed, i.e., during inference and usage.

The Role of XAI. Where formal implementation of fairness during model development is difficult or human oversight is required, XAI is crucial to inform human discretion such that fairness objectives can still be realized. In this sense, XAI is commonly proposed to support human decision-makers and domain experts in fostering fairer decisions (e.g., [100, 126, 130]). For example, if problematic behavioral patterns of users are observed, XAI could provide tailored information to reconsider decisions, similar to Miller's [97] idea of "evaluative AI". Beyond simple recommendations, such information may include comparison to similar instances [25], disclosure of uncertainty [16], or conditional heatmaps in computer vision tasks [2]. However, both designing XAI and training human decision-makers to interact with the explanations is challenging [81]. As feature importance explanations may even hinder fairness of human decisions [120], we are in need of more conceptual and empirical research on how to design XAI towards fairness objectives in human-in-the-loop settings (e.g., how to effectively override certain types of AI recommendations that violate fairness objectives).

3.6 Empowering Fairness

> *"From the perspective of individuals affected by automated decision-making, we propose three aims for explanations: (1) to inform and help the individual understand why a particular decision was reached, (2) to provide grounds to contest the decision if the outcome is undesired, and (3) to understand what could be changed to receive a desired result in the future, based on the current decision-making model."* [134, p. 2]

Fairness considerations do not end after an AI model has been designed and decisions have been made. *Empowering fairness* refers to the ability of affected parties to take effective actions regarding their outcome of a particular decision, e.g., by contesting decisions or seeking recourse.

In their article on the right to explanation, Vredenburgh [133] proposes two types of *self-informed advocacy*: retrospectively, affected parties should be able to identify the accountable entity to demand remedy for unfair treatment (viz., responsibility); prospectively, contesting and recourse options should enable affected parties to actively improve upon their possibly unfair outcome (viz., agency). Our conception of empowering fairness strongly relates to Vredenburgh's forward-looking self-informed advocacy. It more generally relates to attempts of conceptualizing what makes a *fair explanation* in the context of the right to explanation [63].

Stage of the Lifecycle. Recourse and contesting is only possible after a decision has been made. Accordingly, we map empowering fairness to inference & usage.

The Role of XAI. XAI could be useful for empowering fairness because it may help acquire the information to be communicated. Counterfactual explanations have been claimed to provide a promising solution to inform and empower affected parties [10, 134]. For example, they can clarify what combination of feature values would lead to a different outcome. Therefore, counterfactual explanations can provide valuable information on recourse, e.g., by telling a loan applicant how much more she would need to earn in order to be granted a loan [134]. While research on counterfactual explanations is growing rapidly, they do not come without limitations regarding their validity [59], underlying assumptions [68], or susceptibility to manipulations [127]. Extending the scope beyond model-centered explanations, XAI might also involve practical information about responsible contact persons, guidance on how to seek redress, or collaborative platforms for affected parties to share experienced outcomes [80].

3.7 Long-Term Fairness

> *"Accuracy, discrimination, and security characteristics of a system can change over time as well. Simply testing for these problems at training time [...] is not adequate for high-stakes, human-centered, or regulated ML systems. Accuracy, discrimination, and security should be monitored in real-time and over time, as long as a model is deployed."* [55, p. 18]

Fairness remains relevant over the entire AI lifecycle, even after deployment. Hence, *long-term fairness* captures the dynamic interplay of an AI system with the socio-technical system it is deployed in over time.

The long-term impact of an AI model can be analyzed from several perspectives. Arif Khan et al. [9] contrast "formal" and "substantive" equality of opportunity where a forward-facing view of fair life chances also accounts for affected parties' future prospects of success (as opposed to ensuring fair contests at a discrete point in time). Since such conceptions make assumptions about structural disadvantages and future prospects of affected parties, they require much broader and longitudinal evaluation. Further, due to concept drift, a model's formal fairness properties can change over time and should be monitored.

Stage of the Lifecycle. We map long-term fairness primarily onto usage & inference, noting that it may encompass all future iterations of the AI lifecycle until the AI system is shut down.

The Role of XAI. Monitoring tools may help to track changes of fairness metrics and identify situations where interventions are necessary [24, 53, 106]. Another long-term fairness impact is strategic gaming behavior that arises from transparent models [87]. For example, loan applicants who receive counterfactual explanations exactly describing how to reach the decision threshold are prone to create a game-theoretic situation where information itself can be unfairly distributed among clients. We suspect several other forms of long-term fairness issues to emerge that are currently under-explored in the literature. For example, Liu et al. [92] model the impact of formal fairness on the underlying population over time, which Hardt et al. [66] coined as "performative power". Novel forms of XAI may help to anticipate and evaluate such dynamics.

3.8 Informational Fairness

> "*Transparency [...] is valuable because and in so far as it enables the individuals, who are subjected to algorithmic decision-making, to assess whether these decisions are morally and politically justifiable.*" [93, p. 254]

All of the listed fairness desiderata can be augmented with a meta desideratum targeted at transparency *about* fairness, which we label as *informational fairness*.

Informational fairness was originally introduced as a psychological construct in the context of organizational justice [30] to test whether the communication accompanying a decision is candid, truthful, reasonable, timely, and specific. Our conception of informational fairness is inspired by this construct and corresponds to a great extent to Loi et al. [93]'s concept of design publicity. In line with our idea of broad XAI, Loi et al. [93] demand a form of transparency that not only explains a model and its underlying functioning but also the goals and values that went into its design and how these are embedded in the model. A similar distinction is made by Walmsley [136] who differentiates between functional transparency concerned with the inner workings of an AI model and outward transparency related to the communication with stakeholders.

Stage of the Lifecycle. We conceive informational fairness as a meta desideratum that applies to all other fairness desiderata (see Figure 2). Accordingly, informational fairness can be considered across all stages of the AI lifecycle from problem formulation to inference & usage.

The Role of XAI. In the case of informational fairness, XAI is not an aid to but the desideratum itself. Following this idea, we provide examples of what could be made transparent and communicated to affected parties for each fairness desideratum. Regarding fairness understanding, developers and deployers could justify their fairness objectives, delineate potential tradeoffs, and elucidate the mechanisms through which AI aims to achieve them [93]. For data fairness, they could explain how the training data is composed highlighting potential sources of bias [8]. Explaining the training objectives and outlining potential uncertainties or goal conflicts might be appropriate to address formal fairness [98]. Regarding perceived fairness, explaining the functions and limitations of XAI-generated information could enhance understanding to approach "appropriate fairness perceptions" [121]. Transparency about the role of AI and the responsibilities of a human decision-maker within the decision-making context might be desirable for fairness with human oversight [95]. The idea of empowering fairness relies on truthful, understandable, and actionable information [138]. Lastly, regarding long-term fairness, unexpected downstream impacts and the factors driving potentially unfair dynamics can be communicated to stakeholders [126].

4 THE COMPAS CASE

To illustrate the utility of our mapping, we describe how XAI could help address fairness throughout the lifecycle of a high-stake AI system. As an example, we consider the recidivism prediction software COMPAS developed by Northpointe (today rebranded to *equivant Supervision*). COMPAS has been installed in many US-American jurisdictions in order to predict whether a defendant will likely commit another crime in the near future. Judges use this information, e.g., for decisions about who they release on bail. However, COMPAS has been criticized by investigative journalists at ProPublica for disadvantaging Black people [7]. This is especially problematic given the history of systematic discrimination and marginalization of Black people in the US [108]. In this section, we illustrate both how Northpointe could have addressed different fairness desiderata during system development and how they could still, given COMPAS' continued use, address some of them. Of course, XAI is not the only means to this end and is not to be conceived as an ethical panacea. Notably, our illustration also presupposes an inherent motivation to actually address fairness desiderata, which is not necessarily given in the context of a profit-oriented proprietary system like COMPAS.

Let us proceed along the AI lifecycle. First, Northpointe could have used XAI to ensure that the development is based on an appropriate *fairness understanding*. ProPublica's analysis of COMPAS shows that although it was tested for some fairness objective (viz., predictive parity), it falls short on other fairness objectives (viz., equalized odds) in problematic ways [29]. Thus, predictive parity may have been an inadequate fairness objective. Northpointe could also have based their development on insights about the predictive value and correlations of certain features. It has been shown that

defendants' age and number of previous crimes are most predictive for recidivism [40, 115] but also highly correlated to race. Based on such insights, Northpointe could have taken a position on how to treat such features considering pre-existing systematic discrimination.

To address *data fairness*, Northpointe would have needed to ensure that the data set adequately represents demographic groups targeted by the system and to be aware of existing structural relationships such as statistically higher crime rates in predominantly Black neighborhoods [17]. Descriptive statistics could have already pointed them towards biases introduced in the data collection process—a reason to reiterate the data collection stage to apply new data collection strategies [88]. Statistical analysis could also have helped unveil traces of systematic discrimination in the data (e.g. that Black people are more likely to be arrested for minor offenses due to increased police presence in Black neighborhoods). Northpointe could have used such insights either to change their data collection strategies (e.g., collecting features that are less correlated to race), or they could have used them during the feature selection phase (e.g., to mitigate systematic discrimination at the data-level [109]).

Regarding *formal fairness*, Northpointe tested COMPAS for equality of error rates (viz., predictive parity). By contrast, journalists at ProPublica tested for the balance of true and false positives (viz., equalized odds) [29]. Selecting appropriate formal fairness metrics is an intricate endeavor [35]. However, had Northpointe tested COMPAS for equalized odds, too, they might have prevented substantial societal harms. Today, developers can draw on an extensive suite of fairness testing tools for a comprehensive formal fairness assessment [27, 36, 90].

Northpointe could have also detected flaws by engaging with stakeholders and gathering feedback on *perceived fairness*. Outcome statistics, fairness metrics, or model explanations (e.g., in the form of counterfactuals) could have been presented to a focus group of diverse stakeholders to assess and discuss fairness perceptions (cf. [130]). Potential concerns could be handled by reconciling the different perspectives, and even today (after deployment) Northpointe could continue to monitor fairness perceptions by providing feedback channels.

Our discussion thus far has primarily focused on measures that COMPAS' developers could have taken early in the AI lifecycle. But even now, during COMPAS' continued usage, Northpointe could still address several fairness issues via XAI. First, they could seek *fairness with human oversight*. This is a particularly relevant desideratum because COMPAS does not make autonomous decisions but informs the decisions of judges. So, Northpointe could familiarize judges with the basic functioning, strengths, limitations, and uncertainties of the system as well as the factors driving the risk score to ensure judges will use it appropriately. Further, analyzing or supervising the judges' interactions with the system and its explanations might help to achieve fairness objectives more effectively [131].

Northpointe could also foster *empowering fairness* by providing defendants with two kinds of explanations: First, global explanations [39] about the general functioning of COMPAS could be communicated. Where relevant factors are under a person's control (as, e.g., with defendants' history of misdemeanor or substance abuse), insights about their impact can allow people to make life choices decreasing their risk scores. Second, local explanations [39] could provide information about an individual's risk score granting the opportunity to (justifiably) contest the score, e.g., by denouncing discriminatory treatment [70], or to (effectively) seek recourse, e.g., by signaling plans to go into rehab [134].

Moreover, Northpointe could monitor *long-term fairness* by tracking the specified formal fairness objective(s) over time and accordingly adjust the model over several iterations of the AI lifecycle [24]. Further, changes in defendants' behavior due to specified fairness objectives or increased transparency could be examined. For example, precise global or counterfactual explanations might allow defendants to strategically game the system in unexpected ways [87], which might—similar to unfair recourse [61]—be easier for some than for others.

Finally, Northpointe could have aimed, and still could aim, to ensure *informational fairness*. To this end, they would have needed to document fairness-related information and communicate it accordingly. For example, Northpointe could have used model (fairness) cards similar to [98], which would have clarified many of the issues that were only revealed due to the investigative work of Angwin et al. [7]. Most importantly perhaps, Northpointe could have been candid about the underlying business and fairness objectives from the start, ideally making transparent decisions about tradeoffs (e.g., when looking at predictive parity rather than equalized odds) and justifying why an AI system was an appropriate tool for criminal risk prediction in the first place. In communication with affected parties, COMPAS could be complemented with truthful, understandable, and relevant information [10] explaining the underlying logic. At the same time, jurisdictions could be more transparent about how exactly COMPAS is employed and how judges interact with the system. Overall, the strategic use of XAI along the AI lifecycle could benefit all fairness desiderata regarding a model like COMPAS—given sincere intentions to actually pursue these desiderata.

5 CONCLUSION AND OUTLOOK

We distilled eight fairness desiderata from interdisciplinary literature, mapped them onto the AI lifecycle, and discussed how XAI might be leveraged to contribute to fulfilling them. Finally, we illustrated the utility of our approach by applying it to the COMPAS case. The overall picture we paint (see Figure 2) highlights that to effectively design and utilize XAI for fairness, it is paramount to reflect about which fairness desideratum one seeks to fulfill. Before closing, we shall briefly comment on some limitations of our proposal and point to avenues for future work.

Conceptualization and Validation. We do not claim our proposed fairness desiderata to be mutually exclusive or collectively exhaustive. As our discussion in Section 3 has shown, different terms elicit different associations in different communities and conceptual overlap between different fairness desiderata seems difficult to avoid altogether (e.g., long-term fairness to some extent includes formal fairness, etc.). Yet, we are confident that our application to the COMPAS case highlights the general usefulness of our work. Thus, we firmly believe that our eight fairness desiderata provide a valuable starting point for refinement in follow-up work. We expect that future interdisciplinary research can—through both conceptual work and validation by application to real-world cases—provide more sharply distinguished categories, capture more fine-grained distinctions, and incorporate an even broader spectrum of perspectives on fairness. Specific open issues include, e.g., where to locate procedural fairness in our account. For now, we deliberately excluded that notion as it has inherently differing meanings across (and sometimes even within) disciplines. Another open question concerns the status of privacy. Legal scholars might consider data privacy a form of fairness; conceived this way, privacy might qualify as a form data fairness.

Generalizability. Our discussion has focused primarily on high-risk applications and more traditional decision-support systems. To what extent does our proposal generalize to low-risk applications? And what about generative AI? Consider Google's multi-modal AI-chatbot Gemini, which has recently received bad publicity for its inaccurate historical depictions—which were driven by misdirected fairness interventions [1]. We suspect that the usage of (seemingly low-risk) AI systems at scale will accumulate to significant societal impacts bearing more subtle threats to fairness such as, e.g., representational harms [12] in the Gemini case. Thus, although contemporary regulation (like the European AI Act) focuses predominantly on high-risk systems, fairness is an important concern for low-risk applications, too. Beyond that, the Gemini case illustrates that fairness considerations affect generative AI just as much as more traditional AI systems for decision-making. Given the rapid advancement and broad adoption of generative AI, we believe that fairness considerations will be indispensable in this context, though they only start to gain traction [50, 101]. What makes them particularly challenging is that traditional stakeholder roles may no longer hold: in the Gemini case, users

and affected parties are the same people; and ChatGPT users creating personalized chatbots are no longer "just users" either. Thus, to ensure fairness, future research might not only have to refine fairness desiderata but also rethink stakeholder categories.

Actionability. We acknowledge that our proposal does not present an actionable process model for fair software engineering; nor does it provide regulatory guidelines to be enforced by legal institutions. Still, it may serve as a useful roadmap for researchers, developers, and regulators. For researchers, it provides a starting point for a truly interdisciplinary discourse going beyond discussions of algorithmic fairness focusing on technical aspects of AI systems. For developers, our mapping provides urgently needed guidance as to what fairness challenges should be addressed when developing specific guidelines and requirements (see Hacker [62]). More specifically, it may help determine under which circumstances human oversight is beneficial [120, 131] or how responsibilities can be attributed along the AI lifecycle [112]. Naturally, the implications of our approach will vary with the precise settings; so fine-tuning may be needed for different business, societal, and legal contexts.

REFERENCES

[1] 2024. Google pauses AI-generated images of people after ethnicity criticism. *The Guardian* (2024). https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

[2] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence* 5, 9 (2023), 1006–1019.

[3] Avinash Agarwal and Harsh Agarwal. 2023. A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. *AI and Ethics* (2023), 1–16.

[4] Yongsu Ahn and Yu-Ru Lin. 2020. FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1086–1095.

[5] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: The risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning,*. 161–170.

[6] Andrea Aler Tubella, Dimitri Coelho Mollo, Adam Dahlgren Lindström, Hannah Devinney, Virginia Dignum, Petter Ericson, Anna Jonsson, Timotheus Kampik, Tom Lenaerts, Julian Alfredo Mendez, and Juan Carlos Nieves. 2023. ACROCPoLis: A Descriptive Framework for Making Sense of Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library)*. Association for Computing Machinery, 1014–1025.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. In *Ethics of data and analytics*, Kirsten Martin (Ed.). CRC Press Taylor & Francis Group, Boca Raton and London and New York, 254–264.

[8] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

[9] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization (ACM Digital Library)*. Association for Computing Machinery, 1–10.

[10] Nicholas Asher, Lucas de Lara, Soumya Paul, and Chris Russell. 2022. Counterfactual Models for Fair and Adequate Explanations. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 316–349.

[11] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2022. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*.

[12] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society (SIGCIS)*.

[13] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

[14] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* (2016).

[15] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 1–72.

[16] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*

(ACM Digital Library). Association for Computing Machinery, 401–413.

[17] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency* 81 (2018), 149–159.

[18] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM.

[19] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage" Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI*. ACM, 1–14.

[20] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. FlipTest: Fairness Testing via Optimal Transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 111–121.

[21] William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. 2020. Fair Allocation through Selective Information Acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (ACM Digital Library)*. Association for Computing Machinery, 22–28.

[22] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8, Article 832 (2019), 34 pages. https://doi.org/10.3390/electronics8080832

[23] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.

[24] Alessandro Castelnovo, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Cosentini. 2021. Towards Fairness Through Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 647–663.

[25] Joymallya Chakraborty, Kewen Peng, and Tim Menzies. 2020. Making fair ML software using trustworthy explanation. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. ACM, 1229–1233.

[26] Larissa Chazette, Wasja Brunotte, and Timo Speith. 2021. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *Proceedings of the 29th IEEE International Requirements Engineering Conference*. IEEE, Piscataway, NJ, USA, 197–208. https://doi.org/10.1109/RE51729.2021.00025

[27] Zhenpeng Chen, Jie M Zhang, Max Hort, Mark Harman, and Federica Sarro. 2023. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Transactions on Software Engineering and Methodology* (2023).

[28] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[29] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.

[30] J. A. Colquitt. 2001. On the dimensionality of organizational justice: a construct validation of a measure. *The Journal of applied psychology* 86, 3 (2001), 386–400.

[31] Jason A. Colquitt and Jessica B. Rodell. 2015. Measuring Justice and Fairness. In *The Oxford handbook of justice in the workplace*, Russell Cropanzano, Russell S. Cropanzano, and Maureen L. Ambrose (Eds.). Oxford University Press, Oxford.

[32] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. http://arxiv.org/pdf/1808.00023v2

[33] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241–1299. https://doi.org/10.2307/1229039 Publisher: Stanford Law Review.

[34] Jeffrey Dastin. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women *. In *Ethics of data and analytics*, Kirsten Martin (Ed.). CRC Press Taylor & Francis Group, Boca Raton and London and New York, 296–299.

[35] Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. 2022. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management* 31, 10 (2022), 3749–3770.

[36] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2023. A Critical Survey on Fairness Benefits of XAI. http://arxiv.org/pdf/2310.13007.pdf

[37] Marybeth Defrance and Tijl de Bie. 2023. Maximal fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library)*. Association for Computing Machinery, 851–880.

[38] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 1591–1602.

[39] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[40] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018).

[41] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.

[42] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. ACM Press.

[43] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (ACM Digital Library)*. Association for Computing Machinery, 1–6.

[44] Lena Enqvist. 2023. 'Human oversight' in the EU artificial intelligence act: what, when and by whom? *Law, Innovation and Technology* 15, 2 (2023), 508–535.

[45] European Commission. Directorate General for Communications Networks, Content and Technology. and High Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[46] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. 2022. Explanation-Guided Fairness Testing through Genetic Algorithm. In *Proceedings of the 44th International Conference on Software Engineering (ACM Digital Library)*. Association for Computing Machinery, 871–882.

[47] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37–37.

[48] Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16, 8 (2021), e12760.

[49] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness. *Communications of the ACM* 64, 4 (2021), 136–143.

[50] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. http://arxiv.org/pdf/2302.10893

[51] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadoral, and Ansgar Walther. 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance* 77, 1 (2022), 5–47.

[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[53] Avijit Ghosh, Aalok Shanbhag, and Christo Wilson. 2022. FairCanary: Rapid Continuous Explainable Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22)*.

[54] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. 2023. "How Biased is Your Feature?": Computing Fairness Influence Functions with Global Sensitivity Analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library)*. Association for Computing Machinery.

[55] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. 2020. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information* 11, 3 (2020), 1–32.

[56] Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. 2019. Explaining Explanations to Society. In *Proceedings of the NeurIPS 2018 Workshop on Ethical, Social and Governance Issues in AI* (Montréal, Québec, Canada), Chloé Bakalar, Sarah Bird, Tiberio Caetano, Edward Felten, Dario Garcia-Garcia, Isabel Kloumann, Finn Lattimore, Sendhil Mullainathan, and D. Sculley (Eds.). 1–6. arXiv:1901.06560

[57] Przemyslaw Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM.

[58] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (2022).

[59] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55.

[60] Soumyajit Gupta, Sooyong Lee, Maria De-Arteaga, and Matthew Lease. 2023. Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection. In *Proceedings of the ACM Web Conference 2023*. Association for Computing Machinery and Saarländische Universitäts- und Landesbibliothek, 3689–3700.

[61] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. arxiv preprint. https://arxiv.org/abs/1909.03166

[62] Philipp Hacker. 2022. The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279796

[63] Philipp Hacker and Jan-Hendrik Passoth. 2022. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In *xxAI – Beyond Explainable AI*, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (Eds.). Springer eBook Collection, Vol. 13200. Springer International Publishing and Imprint Springer, Cham, 343–373.

[64] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[65] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, Ashish Rathi, Scott Rees, Ankit Siva, ErhYuan Tsai, Keerthan Vasist, Pinar Yilmaz, Muhammad Bilal Zafar, Sanjiv Das, Kevin Haas, Tyler Hill, and Krishnaram Kenthapadi. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021)*.

[66] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. 2022. Performative Power. *Advances in Neural Information Processing Systems* 35 (2022), 22969–22981.

[67] Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

[68] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 513.

[69] John E. Hunter, Frank L. Schmidt, and Ronda Hunter. 1979. Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin* 86, 4 (1979), 721–735.

[70] Jean-Marie John-Mathews. 2022. Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change* 174 (2022), 1–29.

[71] Arif Ali Khan, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. 2022. Ethics of AI: A Systematic Literature Review of Principles and Challenges. In *The International Conference on Evaluation and Assessment in Software Engineering 2022 (ACM Digital Library)*. Association for Computing Machinery, 383–392.

[72] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores: 67(43): 1–23. *Leibniz International Proceedings in Informatics* 67, 43 (2017), 1–23.

[73] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. 2019. Explainability as a Non-Functional Requirement. In *Proceedings of the 27th IEEE International Requirements Engineering Conference*, Daniela E. Damian, Anna Perini, and Seok-Won Lee (Eds.). IEEE, Piscataway, NJ, USA, 363–368. https://doi.org/10.1109/RE.2019.00046

[74] Niklas Kühl, Robin Hirt, Lucas Baier, Björn Schmitz, and Gerhard Satzger. 2021. How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. *Communications of the Association for Information Systems* 48, 1 (2021), 46.

[75] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[76] Markus Langer. 2023. Fehlgeleitete Hoffnungen? Grenzen menschlicher Aufsicht beim Einsatz algorithmusbasierter Systeme am Beispiel Personalauswahl. *Psychologische Rundschau* 74, 4 (2023), 211–220.

[77] Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, and Jonas Wahl. 2021. Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives. In *29th IEEE International Requirements Engineering Conference Workshops* (Notre Dame, Indiana, USA) *(REW 2021)*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, Piscataway, NJ, USA, 164–168. https://doi.org/10.1109/REW53955.2021.00030

[78] Markus Langer and Richard N. Landers. 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 (2021), 106878.

[79] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 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 (2021), 1–24.

[80] Erwan Le Merrer and Gilles Trédan. 2020. Remote explainability faces the bouncer problem. *Nature Machine Intelligence* 2, 9 (2020), 529–539.

[81] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 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 (2022), 126–148.

[82] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 205395171875668.

[83] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–26.

[84] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[85] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics* 1, 4 (Nov. 2021), 529–544. https://doi.org/10.1007/s43681-021-00067-y

[86] Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 704–714.

[87] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[88] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (ACM Digital Library)*. Association for Computing Machinery, 396–410.

[89] Yi Li and Nuno Vasconcelos. 2019. REPAIR: Removing Representation Bias by Dataset Resampling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[90] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (2020), 1–45.

[91] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 8136–8146.

[92] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. *International Conference on Machine Learning* (2018), 3150–3158.

[93] Michele Loi, Andrea Ferrario, and Eleonora Viganò. 2021. Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and information technology* 23, 3 (2021), 253–263.

[94] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems*.

[95] Kirsten Martin. 2019. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 160, 4 (2019), 835–850.

[96] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computer Surveys* 54, 6 (2021).

[97] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI!. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library)*. Association for Computing Machinery, 333–342.

[98] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (ACM Conferences)*. ACM, 220–229.

[99] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[100] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. *ACM Transactions on Interactive Intelligent Systems* 12, 3 (2022), 1–30.

[101] Terrence Neumann, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. 2024. Diverse, but Divisive: LLMs Can Exaggerate Gender Differences in Opinion Related to Harms of Misinformation. *arXiv preprint arXiv:2401.16558* (2024).

[102] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A Voting-Based System for Ethical Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018).

[103] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (2020).

[104] Julian Nyarko, Sharad Goel, and Roseanna Sommers. 2021. Breaking Taboos in Fair Machine Learning: An Experimental Study. In *Equity and Access in Algorithms, Mechanisms, and Optimization (ACM Digital Library)*. Association for Computing Machinery, 1–11.

[105] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29, 3 (2019), 441–459. https://doi.org/10.1007/s11023-019-09502-w

[106] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 2021. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management* 58, 5 (2021), 1–17.

[107] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In *Proceedings of the IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 94–100.

[108] Becky Pettit and Carmen Gutierrez. 2018. Mass Incarceration and Racial Inequality. *American journal of economics and sociology* 77, 3-4 (2018), 1153–1182. https://doi.org/10.1111/ajes.12241

[109] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable Data-Based Explanations for Fairness Debugging. In *Proceedings of the 2022 International Conference on Management of Data (ACM Digital Library)*. Association for Computing Machinery, 247–261.

[110] Alexandre Quemy. 2020. Two-stage optimization for machine learning workflow. *Information Systems* 92 (2020), 101483.

[111] Iyad Rahwan. 2018. Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology* 20, 1 (2018).

[112] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 33–44.

[113] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, Cham.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

[115] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[116] Pedro Saleiro, Kit T. Rodolfa, and Rayid Ghani. 2020. Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (ACM Digital Library)*. Association for Computing Machinery, 3513–3514.

[117] Laura Sartori and Andreas Theodorou. 2022. A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control. *Ethics and Information Technology* 24, 1 (2022), 1–11.

[118] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[119] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library)*. Association for Computing Machinery, 959–970.

[120] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[121] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems.. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

[122] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM.

[123] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. 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* (2022), 1–28.

[124] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 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 (2022), 1–13.

[125] Vivek Singh, Anshuman Singh, and Kailash Joshi. 2022. Fair CRISP-DM: Embedding fairness in machine learning (ML) development life cycle. (2022).

[126] Dylan Slack, Sorelle A. Friedler, and Emile Givental. 2020. Fairness warnings and Fair-MAML: Learning Fairly with Minimal Data. In *Conference on Fairness, Accountability, and Transparency (FAT '20)*.

[127] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc, 62–75.

[128] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 1–16.

[129] Georg Starke, Benedikt Schmidt, Eva de Clercq, and Bernice Simone Elger. 2022. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI and Ethics* (2022), 1–12.

[130] Simone Stumpf, Lorenzo Strappelli, Subeida Ahmed, Yuri Nakao, Aisha Naseer, Giulia Del Gamba, and Daniele Regoli. 2021. Design Methods for Artificial Intelligence Fairness and Transparency. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[131] Mike Teodorescu, Lily Morse, Yazeed Awwad, and Gerald Kane. 2021. Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *Management Information Systems Quarterly* 45, 3 (2021), 1483–1500.

[132] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (ACM Conferences)*. ACM, 1–7.

[133] Kate Vredenburgh. 2022. The Right to Explanation*. *Journal of Political Philosophy* 30, 2 (2022), 209–229.

[134] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017), 1–47.

[135] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AIQ1BComputer Law & Security Review; H-Index: 41 SJR: Q1 CORE: NA ABDC: B FT50: NA C2Computer Law & Security Review; H-Index: 41 VHB: NA FNEGE: NA CoNRS: NA HCERE: NA CCF: C BFI: 2 AJG: NA +. *Computer Law & Security Review* 41 (2021).

[136] Joel Walmsley. 2021. Artificial intelligence and the value of transparency. *AI & SOCIETY* 36, 2 (2021), 585–595.

[137] Mowei Wang, Yong Cui, Xin Wang, Shihan Xiao, and Junchen Jiang. 2017. Machine learning for networking: Workflow, advances and opportunities. *Ieee Network* 32, 2 (2017), 92–99.

[138] David S. Watson and Luciano Floridi. 2021. The Explanation Game: A Formal Framework for Interpretable Machine Learning. In *Ethics, Governance, and Policies in Artificial Intelligence*, Luciano Floridi (Ed.). Springer eBook Collection, Vol. 144. Springer International Publishing and Imprint Springer, Cham.

[139] Rüdiger Wirth and Jochen Hipp. 2000. Crisp-dm: towards a standard process modell for data mining. https://api.semanticscholar.org/CorpusID:1211505

[140] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. 2021. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (ACM Digital Library)*. Association for Computing Machinery, 1023–1033.