A Critical Survey on Fairness Benefits of Explainable AI

Luca Deck
luca.deck@uni-bayreuth.de
University of Bayreuth & Fraunhofer FIT
Bayreuth, Germany

Maria De-Arteaga
dearteaga@utexas.edu
University of Texas at Austin
Austin, USA

Jakob Schoeffer
schoeffer@utexas.edu
University of Texas at Austin
Austin, USA

Niklas Kühl
kuehl@uni-bayreuth.de
University of Bayreuth & Fraunhofer FIT
Bayreuth, Germany

ABSTRACT

In this critical survey, we analyze typical claims on the relationship between explainable AI (XAI) and fairness to disentangle the multidimensional relationship between these two concepts. Based on a systematic literature review and a subsequent qualitative content analysis, we identify seven archetypal claims from 175 scientific articles on the alleged fairness benefits of XAI. We present crucial caveats with respect to these claims and provide an entry point for future discussions around the potentials and limitations of XAI for specific fairness desiderata. Importantly, we notice that claims are often (*i*) vague and simplistic, (*ii*) lacking normative grounding, or (*iii*) poorly aligned with the actual capabilities of XAI. We suggest to conceive XAI not as an ethical panacea but as one of many tools to approach the multidimensional, sociotechnical challenge of algorithmic fairness. Moreover, when making a claim about XAI and fairness, we emphasize the need to be more specific about *what* kind of XAI method is used, *which* fairness desideratum it refers to, *how* exactly it enables fairness, and *who* is the stakeholder that benefits from XAI.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Information systems** → **Decision support systems**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Explainable AI, algorithmic fairness, critical survey

1 INTRODUCTION

The integration of artificial intelligence (AI) into decision-making processes has raised concerns about reinforcing societal inequalities [17, 151]. Moreover, much progress in AI comes at the cost of increased complexity and opacity, which may impede human understanding [38]. Explainable AI (XAI) is commonly conceived as a remedy to both of these challenges [20]. However, the implicit link between XAI and fairness has been challenged due to inconclusive evidence and a lack of consistent terminology [23, 122].

Our critical survey explores the complex relationship between XAI and algorithmic fairness by reviewing 175 recent articles and identifying seven archetypal claims on the alleged fairness benefits of XAI. Organizing the scattered debate into meaningful subdebates, we discuss caveats and provide an entry point for future discussions on the suitability and limitations of XAI for fairness. We find that literature from various domains is highly optimistic about the usefulness of XAI for several fairness dimensions and stakeholders. However, many claims in the literature remain vague about how exactly XAI will contribute to fairness. They disregard technical limitations, conflicts of interest between stakeholders, and normative grounding. Highlighting central caveats as well as nascent approaches to address these caveats, we contribute to a more nuanced understanding of the interplay between XAI and fairness in AI-informed decision-making.

This article is structured as follows: we start by establishing key concepts that we use to structure the debate and interpret claims. Next, we describe how we identified and organized claims from 175 articles into 7 archetypal categories. Afterwards, we introduce each archetypal claim at face value by verbalizing the underlying intuition from the literature. Based on that, we organize a meaningful, structured debate and take a critical perspective on these claims. Finally, we synthesize patterns of critique and caveats to embed the alleged fairness benefits of XAI in a bigger picture. Through this thorough and systematic study, we hope to bring clarity to the entangled relationship between XAI and fairness and inform future efforts aimed at leveraging XAI to tackle algorithmic unfairness.

2 BACKGROUND

Related work. XAI and fairness have both separately produced a large body of research [45, 202]. Recent work has noted that while claims regarding the intersection are frequent, there is a lack of specificity regarding how they relate [23, 140]. Langer et al. [122] extract different desiderata of XAI (with one of them being fairness)

from over 100 peer-reviewed publications and point out that only a subset of articles substantiates their claims with empirical evidence. Balkir et al. [23] survey challenges in the application of XAI for fairer language models and call for more precise conceptualization on how exactly XAI relates to fairness. Zhou et al. [237] provide an overview of XAI methods for fairness objectives and highlight the consideration of contextual factors as well as the need for interdisciplinary research. Our work adds to this stream of literature by adopting a systematic and critical approach similar to Blodgett et al. [36], who critically analyze the use of the term "bias" in the context of natural language processing.

Dimensions of XAI and fairness. Both XAI and fairness are multifaceted concepts that can be conceptualized along various dimensions [45, 202]. In this work, we understand XAI as any mechanism that "produces details or reasons to make [the] functioning [of an AI system] clear or easy to understand" [20]. We adopt the term *desideratum* from Langer et al. [122] to differentiate several roles and objectives of XAI. We explicitly examine desiderata in the context of fairness and distinguish between formalized notions of fairness (*formal fairness*) and human perceptions of fairness (*perceived fairness*), similar to Starke et al. [205]. Formal fairness criteria are captured in mathematical and statistical frameworks [24, 42], which may or may not align with human fairness perceptions [66, 152, 204]. Formal fairness notions are often distinguished into group and individual fairness: group fairness criteria typically require a form of parity between demographic groups, for example, along sensitive features like gender or race [50]. Individual fairness criteria typically demand to treat similar people alike [69]. Perceived fairness is a subjective human attitude that is highly context-sensitive [205] and related to complex moral deliberations [30]. It requires fundamentally different measurements than formal fairness, for example, based on psychological constructs [55]. From Colquitt [55], we also adopt the decomposition into *distributive*, *procedural*, and *informational* dimensions. Distributive fairness is primarily concerned with decision outcomes, whereas procedural fairness refers to the underlying decision-making process. Informational fairness accounts for aspects of the communication accompanying a decision.

Regarding the purpose of XAI, we further differentiate *substantial* from *epistemic* goals [122]. An epistemic goal refers to the capability of humans to *observe* fairness properties of a model (e.g., XAI providing insights into a model's reliance on sensitive features), whereas a substantial goal actively aims to *alter* fairness properties (e.g., mitigating formally unfair model characteristics). This distinction is helpful in understanding the multifaceted role of XAI across many application contexts. For example, an epistemic usage of XAI may be to *inform* about a given fairness desideratum (e.g., group fairness), whereas a substantial usage of XAI aims at directly or indirectly *affecting* (un)fairness properties of an AI system. We also adopt the taxonomy from Langer et al. [122] on human stakeholders of XAI, which includes developers, deployers, regulators, users, and affected parties (i.e., decision subjects).

3 METHODOLOGY

Similar to Blodgett et al. [36], we systematically identified and scrutinized claims of recent publications—in our case, about the fairness

benefits of XAI. We first conducted a structured literature review guided by Kitchenham and Charters [112] to identify entrenched claims on the alleged capabilities of XAI for fairness. This process yielded 175 articles and is depicted in Figure 1. We supplemented our deductive literature review with inductive coding [230] at the level of individual claims. A rigorous qualitative analysis of these claims using a grounded theory approach [52] yielded seven archetypal claims, summarized in Figure 2.

3.1 Systematic Literature Review

In order to gain an understanding of the domain, test the effectiveness of keywords, and identify relevant publishers, we initiated an exploratory review by crawling the Google Scholar database. Our search string is chosen to reflect various dimensions of both XAI and fairness and to restrict the results to AI contexts. For XAI, we relied on the taxonomy of Arrieta et al. [20] and also incorporated the related terms *understandability, comprehensibility, interpretability, explainability, and transparency*, as well as their inflections of the same stem. After screening around 400 individual articles, we finally decided on the following search string—note that the asterisk as a wildcard character allows us in each case to consider words of the same stem, including adjectives and nouns:

> (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")

Relying on recent recommendations to combine two popular search strategies, database querying and snowballing [229], we followed proven guidelines for systematic literature reviews in the domain of software engineering [112, 228]. Scopus was the natural choice for our database search because it has been recommended as an effective tool to generate seed sets for snowballing [146], and it includes the most relevant publishers for our task.[1] We allow for snowballing to include publications outside of Scopus using the lens.org database.[2] To account for recent, unreviewed publications, we also applied our search string to the arXiv.org database. Following the documentation guidelines of Kitchenham and Charters [112] and the PRISMA standard [155], we provide a transparent and replicable documentation of the selection process: Figure 1 depicts how a total body of 1,003 identified records (as of September 2022) was condensed to a seed set of 122 with explanations on the filter criteria for each step. Because the goal of this critical survey is to characterize the *contemporary* discourse on XAI and fairness, we focused on articles from 2016 onwards.

At the initial identification level, we only considered scientific articles and excluded records such as courses, keynotes, etc. Here, we manually inspected all abstracts and only retained articles examining dimensions of both XAI *and* fairness fitting into the broader scheme of our definitions. Consequently, we discarded articles having too broad or deviating notions of the XAI terms (e.g., using the term *explain* in a different context), articles using the term *fair* in different contexts (*fairly, FAIR principles*, etc.), and articles where fairness or XAI are not the object of research. Proceeding

---

[1]https://www.elsevier.com/products/scopus
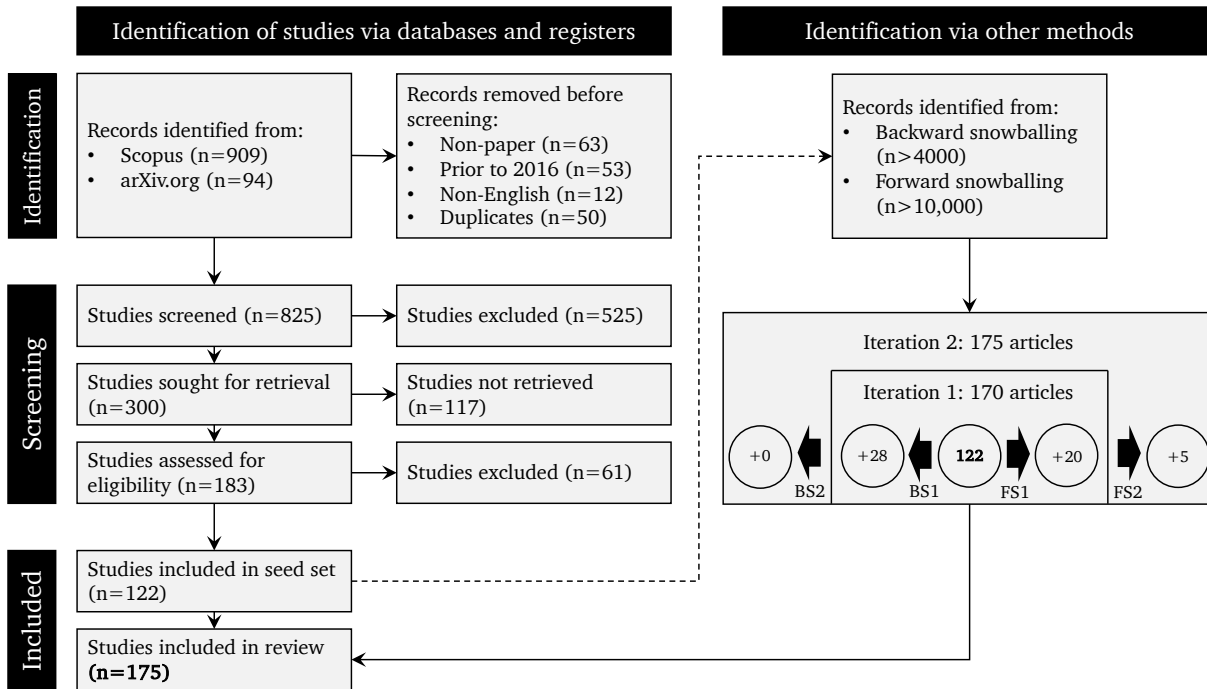[2]https://www.lens.org

Figure 1: PRISMA flowchart describing the article selection procedure.

to full-text analyses, we heuristically scanned the entire article for explicit claims on XAI and fairness. We focused on unique statements as opposed to straightforward summaries or paraphrases of previous work, which eliminated most literature reviews. Finally, we discarded articles where the direct relationship between XAI and fairness was not considered or remained too vague to infer any claims. For example, Shin [189] examines the influence of explainability and fairness on trustworthiness but does not address the interaction of explainability and fairness.

Starting from the seed set of 122 articles, we performed iterative backward and forward snowballing [228]. Using the citation crawling tool Citationchaser [95], we accelerated the snowballing procedure and focused on the most frequently referenced articles. We stopped whenever the third iteration did not generate any further hits. Please refer to Appendix A for a comprehensive overview of the gathered articles categorized according to their underlying methodology.

3.2 Inductive Claim Analysis

In a subsequent step, we inductively identified dominant themes by analyzing commonalities and grouping claims into meaningful categories. We found grounded theory to be an appropriate methodology and followed the research design framework by Chun Tie et al. [52], employing MAXQDA for claim extraction, coding, and memoing [118]. To reiterate, we built our grounded theory around the following two questions:

- What does recent literature claim about the relationship between XAI and fairness?
- On what kind of evidence are these claims grounded?

We started by skimming the full texts of the 175 selected articles to comprehend the respective methodology and key results. In parallel, we scanned for claims with a strong focus on the most relevant and promising article sections. For example, the introduction, discussion, and conclusion sections often provided more expressive claims than the method and results sections, which we usually only considered to retrace methodology or reasoning. Throughout the coding procedure, we used memos to note down important insights, augment the claims with contextual information (such as textual context, meaning of abbreviations, authors' reasoning, etc.), and document the coders' line of thought.

In the first iteration, we kept the codes as specific as possible to maintain a maximum amount of information. During coding, we considered not only the verbatim content of the claims but also their context and, if possible, their underlying reasoning. We used this information to categorize the type of evidence leading to the claim, which we recorded in our coding system. In the subsequent iterations, we identified higher-level concepts and started grouping the claims into mutually exclusive categories. To achieve theoretical saturation, we ensured that the identified categories were sufficient and the assignment was plausible and correct by re-checking each claim [177].

4 CRITICAL SURVEY

Figure 2 shows the result of our coding process and highlights the different roles XAI may take to address fairness desiderata. We use the concepts introduced in Section 2 to organize the claims and relate them to one another. This allows us to differentiate between claims about formal or perceived fairness and claims on what we
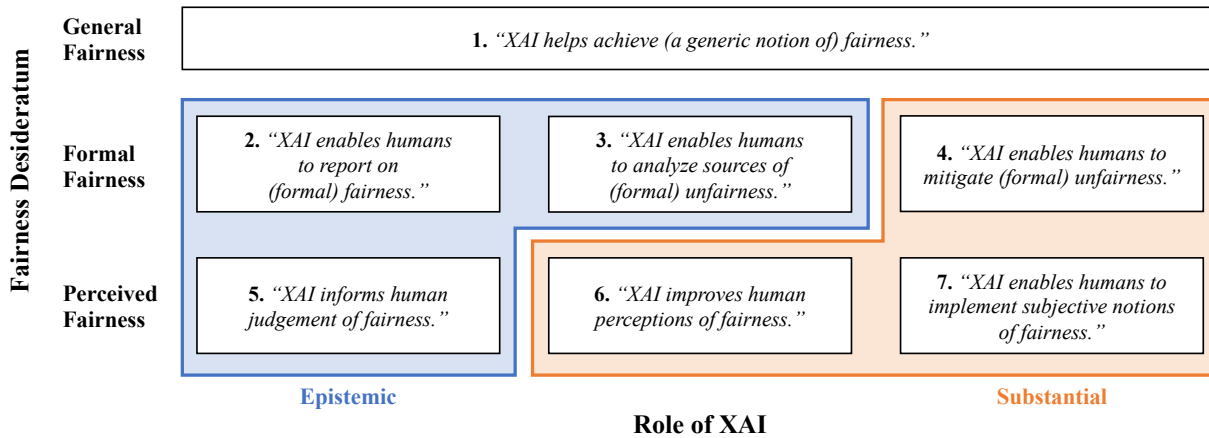
Figure 2: The seven identified archetypal claims on the role of XAI for fairness desiderata.

call "a generic notion of fairness," which do not specify what kind of fairness is pursued. We also differentiate whether the claims relate to an epistemic or a substantial role of XAI. In what follows, we introduce each archetypal claim by verbalizing the underlying intuition and providing representative quotes from the literature. We then organize a structured debate and take a critical perspective. Please refer to Table 1 for a comprehensive overview of archetypal claims, including exemplary supportive and critical quotes from the literature. We also provide a complete list of references per claim in Appendix B.

4.1 Claim 1: *"XAI helps achieve (a generic notion of) fairness."*

This type of claim treats fairness as a monolithic concept without specifying *how* XAI will lead to *which* kind of fairness for *whom*. While phrasing and determinism vary by reference, we identify three tendencies. The first suggests XAI as a *necessary* condition for fairness. For example, a popular belief is that a decision has to be understandable in order to be fair.

> *"First and most evidently, understanding the logic and technical innerworkings (i.e. semantic content) of these systems is a precondition for ensuring their safety and fairness"* [127, p. 40].

XAI is other times treated as *sufficient* for achieving fairness. An exemplary intuition is that revealing the underlying mechanisms of a decision is all it takes to guarantee fairness.

> *"From a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models"* [20, p. 9].

Many others remain *tentative* insofar as they suggest vague capabilities of XAI for fairness but are less assertive.

> *"Explainability and interpretability: these two concepts are seen as possible mechanisms to increase algorithmic fairness, transparency and accountability"* [44, p. 2].

Critique: Futility and danger of vague claims on generic fairness notions. Fairness is a multidimensional concept [55, 147] with many conflicting notions [79, 116]. Hence, it is easy to come up

with counterexamples for such strong and simplistic claims. For example, distributions of classification rates can be used to show that models conform with formal fairness criteria without XAI methods (e.g., [50]). Moreover, a transparent model can be perfectly scrutable and still be deemed unfair by some stakeholders, which precludes the suggested sufficiency [144]. The central underlying assumption behind these claims appears to be that XAI is valuable to *some* dimensions of fairness. This (perhaps plausible) intuition is also reflected in all ethical AI principles reviewed by Floridi et al. [75]. However, we argue that suggesting a universal link between XAI and fairness is misleading and threatens a meaningful debate that should account for the multidimensionality of fairness and incorporate the essential needs of relevant stakeholders [122]. Perpetuating these overly general claims threatens to produce unwarranted trust and reliance on current XAI technologies—and fails to recognize the multitude of interests with regard to fairness. By disentangling the multidimensionality of fairness into more fine-grained scopes, our work organizes the debate in a way that fosters more precise claims and nuanced arguments.

4.2 Claim 2: *"XAI enables humans to report on (formal) fairness."*

A large share of articles is concerned with using XAI to measure and report on formal (un)fairness, often phrased as "identifying bias" [86] or "detecting discrimination" [107]. One central intuition behind these claims is that conventional evaluation of model outcomes (e.g., testing for demographic parity) may fail to consider the underlying mechanisms leading to this outcome [132]. XAI is expected to fill this gap by providing insights into these mechanisms, which may then be related to formal fairness criteria [67]. Since the anti-discriminatory motivation of formal fairness criteria typically relates to sensitive features such as gender or race, XAI is commonly employed to examine how models make use of these sensitive features. Inspired by the legal notion of disparate treatment [131] and the idea of "fairness through unawareness" [120], one form of individual fairness deems any explicit use of sensitive features as unfair (e.g., [12, 97, 106]).

Table 1: Exemplary supportive and critical perspectives on archetypal claims.

Archetypal claim	Stance	Exemplary quotes
"XAI helps achieve (a generic notion of) fairness."	Supportive	"Understanding the logic and technical innerworkings (i.e. semantic content) of these systems is a precondition for ensuring their safety and fairness" [127, p. 40].
	Critical	"[A] perfectly auditable algorithmic decision [...] can nevertheless cause unfair and transformative effects" [144, pp. 14–15].
"XAI enables humans to report on (formal) fairness."	Supportive	"Given a classifier, a dataset and a set of sensitive features, [the proposed method] first assesses whether the classifier is fair by checking its reliance on sensitive features using 'Lime explanations'" [13, p. 1].
	Critical	"We found that despite their capabilities in simplifying and explaining model behavior, many prominent XAI tools lack features that could be critical in detecting bias" [9, p. 1].
"XAI enables humans to analyze sources of (formal) fairness."	Supportive	"To gain insight into how the model discriminates, we construct a transparency report that summarizes the most salient statistical differences in the features between the flipset individuals and their images under the optimal transport mapping [...] Intuitively, the transparency report can serve as an overview of what features the model may be using to discriminate between populations" [35, p. 112].
	Critical	"[LIME] still lacks the skills to detect issues of biased data and detect issues in the selection or processing of the model" [9, p. 12].
"XAI enables humans to mitigate (formal) unfairness."	Supportive	"If deemed unfair, [the proposed XAI-based method] then applies feature dropout to obtain a pool of classifiers. These are then combined into an ensemble classifier that was empirically shown to be less dependent on sensitive features" [13, p. 3].
	Critical	"The best one could do with these [XAI] techniques seem to be to develop 'colourblind' models which, even if they receive explicit information about protected attributes in their input, ignore this information when making their decisions. Although it is simple and intuitive, we suspect that such an approach has similar issues with the much criticized 'fairness through unawareness' approach" [23, p. 8].
"XAI informs human judgement of fairness."	Supportive	"The results [...] suggest that standards clarity and outcome explanation allowed people to judge whether the fairness properties of the algorithm were in line with their fairness concepts" [125, p. 18].
	Critical	"Since the right to explanation as defined in current regulations [...] does not give precise directives on what it means to provide a 'valid explanation' [...], there is a legal loophole that can be used by dishonest companies to cover up the possible unfairness of their black-box models by providing misleading explanations" [6, p. 1].
"XAI improves human perceptions of fairness."	Supportive	"When the system produces a 'positive' output, it may be considered fair and every explanation style (except for demographic-based) works, with a slight advantage to our new [...] explanation" [193, p. 9].
	Critical	"When explanations helped respondents understand biased distributions, perceived fairness decreased" [205, p. 7].
"XAI enables humans to implement subjective notions of fairness."	Supportive	"As a user mentioned, 'Feature selection is sometimes arbitrary, but it provides the feature-level measures as evidence of fairness-aware decision.' — this demonstrated how the system can help decision makers to achieve fair decision making through better explainability" [5, p. 9].
	Critical	"We conjecture that miscalibrated fairness perceptions (e.g., due to misleading explanations) may influence reliance on AI in undesirable ways, by making people adopt incorrect or override correct AI recommendations. This lends support to the hypothesis that there is a disconnect between what explanations provide and the fairness benefits they claim" [181, p. 2].

"An algorithm is fair if the protected features are not explicitly used in the decision-making process" [106, p. 850].

Moreover, in cases where we have clear guidance on which features are legitimate and fair to use, prior work has argued that feature importance can be employed to validate whether models actually rely on these features; for example, in computer vision tasks (e.g., [15]). Prominent XAI methods that have been employed to gauge the use of sensitive features are feature importance measurements using LIME [171] and SHAP [134] (e.g., [13, 46, 107]), or inherently interpretable models (e.g., [140, 169, 211]).

Another common approach to individual fairness focuses on "contrastive explanations" [200] for identifying instances where the outcome changes only based on "flipping" a sensitive feature. Articles differ in their exact methodology but often rely on similarity-based measures (e.g., [35]) or counterfactually constructed data points (e.g., [60]) to identify data pairs where this behavior occurs.

Because sensitive features are often correlated to other (seemingly) legitimate features, some works additionally account for correlations and causal relationships of sensitive features (e.g., [61, 88]).

Critique: Misconceptions of feature importance for group fairness. Logical conclusions between feature inclusion or importance and formal fairness criteria require utmost caution [42, 46]. The idea that only ignoring sensitive features can be sufficient to achieve formal fairness measures, often referred to as "fairness through unawareness" [69], has been shown to be mathematically flawed [114]. Some prior work interprets low feature importance of

sensitive features as a guarantee for demographic parity (e.g., [107]). However, a low feature importance of sensitive features is neither a necessary nor a sufficient criterion to satisfy formal group fairness metrics due to redundant encoding [65, 143] or differing base rates [69, 131]. Redundant encoding (i.e., correlations between sensitive features and other "task-relevant" features) is problematic because many popular post-hoc XAI methods, such as LIME and SHAP, do not account for correlations. Further, when demographic groups have differing base rates (e.g., higher average scores in credit risk scoring), achieving demographic parity would require the decision to account for these group inequalities—which would justify the purposeful use of sensitive features [69].

Critique: Narrow and unacknowledged normative grounding. We further question the normative grounding of using feature inclusion or feature importance to allegedly report on fairness. Prior work has interpreted low feature importance of sensitive features as a form of "fairness through unawareness" (e.g., [106]). This approach not only disregards the possibility of redundant encoding but also makes implicit normative assumptions about the world. As noted by Binns [31], this conception of fairness is based on the assumption that any group differences in features stem from "personal choices," as opposed to structural injustice. These assumptions also become paramount when contrastive explanations are employed to test whether similar individuals are receiving differing outcomes based on their sensitive features (e.g., [35, 72]), because such approaches disregard any relationships between sensitive and other features.

Causality-based and correlation-based XAI methods (e.g., [61, 88]) that aim to resolve the problem of redundant encoding and reveal statistical relationships between sensitive and other features also have implicit normative grounding. In particular, they adopt an ideal mode of theorizing [73], in which they deem something as unjust if it would be unjust in an ideal world. Fazelpour and Lipton [73] have advocated for an inclusion of non-ideal perspectives in our conceptualization of algorithmic fairness, which requires engaging with the ways in which the current world is unfair and acknowledging that what would be fair in an ideal world may not be fair in the face of current injustices. Thus, using the information captured in sensitive features is not always morally wrong [152]. For example, consider the use of standardized tests to predict college success: the financial ability to retake the exam and access training means that the relationship between test scores and target outcomes may vary across demographic groups. Thus, including sensitive features may, in fact, be *desirable* from a fairness perspective [114]. These caveats highlight the fact that a narrow and unacknowledged normative grounding for fairness is pervasive across the literature on XAI. It underlines the need not to misuse XAI as a "fairness proof" but to interpret the fairness reports based on contextual factors and normative deliberations.

Critique: Narrow and unacknowledged views on formal procedural fairness. Balkir et al. [23] observe that XAI and formal fairness criteria take fundamentally different perspectives and argue that XAI tends to focus on the procedural dimension of fairness—that is, one that relates to the fairness of decision-making processes rather than outcomes [89]. Whereas *formal distributive fairness* metrics (such as demographic parity and equality of opportunity)

are well established in the literature and grounded on moral and political philosophy [19, 24], Balkir et al. [23] note that there is less of a shared conception of *formal procedural fairness*. Prior work has often taken the perspective that the use or disuse of sensitive features determines whether an AI system is procedurally fair or not (e.g., [89, 225]). However, this is only a narrow perspective on formal procedural fairness. Morse et al. [145] propose the domain of organizational justice [55, 128] as a source for holistically defining formal procedural fairness. Herein, Leventhal [128] defines six components of procedural fairness: consistency, accuracy, ethicality, representativeness, bias suppression, and correctability. For some of these components (e.g., bias suppression) XAI may in some instances be helpful; others (e.g., ethicality and correctability) would demand measures beyond formal fairness reports, such as value transparency [133] or appeal processes [135]. We discuss this further in Section 4.5.

Critique: Technical limitations of XAI for formal fairness reports. In a plea for intrinsically interpretable models, Rudin [174] argues how model-agnostic explanations of black-box models are fundamentally unfaithful to the original model and cannot explain decision processes sufficiently. This critique has been echoed in multiple studies demonstrating that such approaches are prone to adversarial attacks on fairness reports, which produce innocuous explanations for (formally) unfair models. In fact, a major limitation of feature importance methods is their susceptibility to "fairwashing" through, for example, rationalized surrogate models [6], reliance on input perturbations [199], or exploitation of redundant encoding [65]. Some approaches claim to provide a more "accurate" [84] and "robust" [26] picture of the usage of sensitive features, potentially tackling issues of "fairwashing." However, future work must ascertain the reliability of these analytic tools and be cautious with the interpretation of the analyses.

Rudin [174] further questions the procedural character of many post-hoc XAI methods and argues that they do not reveal direct insights into the true underlying mechanisms of a black-box model. Indeed, post-hoc methods like LIME, SHAP, and contrastive explanations provide little information about the decision-making process itself. Thus, while the goal of these approaches is to report on procedural fairness, a true procedural perspective on fairness may only be provided by model-specific explanations [41], such as intrinsically interpretable rule lists (e.g., [8]).

Critique: Power asymmetries in XAI-enabled fairness reports. The stakeholders (e.g., developers) in charge of producing fairness reports take a crucial role. By making important design choices on the transparency of an AI system, they shape the way a system is perceived by other, mostly less powerful, stakeholders in downstream steps. Many stakeholders without further access and knowledge must rely on the selective information provided by these XAI techniques. This power dynamic is critical, especially since explanations can be manipulative [6, 14, 65, 121]. For XAI to become a valuable tool for fairness desiderata, it is, therefore, important to be explicit about the targeted stakeholders, their potential needs and objectives [122], as well as the normative deliberations that went into the development of relevant XAI techniques [133].

4.3 Claim 3: *"XAI enables humans to analyze sources of (formal) unfairness."*

Beyond descriptive fairness reports, XAI methods are often claimed to uncover patterns of formal unfairness and to pin down contributing factors.

> "The investigations demonstrate that fair decision making requires extensive contextual understanding, and AI explanations help identify potential variables that are driving the unfair outcomes" [237, p. 1].

This extends the epistemic facet of XAI to provide deeper level insights of how a specific notion of (un)fairness emerges. Such claims concern *instance-centric* or *feature-centric* approaches: instance-centric approaches focus on individual instances in the data that drive unfairness. For example, some prior works claim to identify "discriminatory samples" [72] or "discriminatory input" [3] in the training data, which refer to individual instances for which a definition of individual fairness is violated. Feature-centric approaches analyze how features relate to formal fairness. Some extend existing feature importance methods with the goal of quantifying the contribution of features to formal unfairness (e.g., [26, 143]). Others causally decompose the influence of sensitive features on outcomes into direct, indirect, or induced discrimination (e.g., [88, 234]).

Critique: Implicit assumptions when analyzing sources of formal unfairness. Analyzing *sources* of unfairness necessitates a definition of fairness, and as such this body of work suffers from many of the same limitations as approaches aimed to report on (formal) fairness. Unacknowledged normative grounding and misconceptions of feature importance and inclusion, discussed in Section 4.2, are also central weaknesses of existing methodologies that aim to identify sources of unfairness. To avoid repetition, we refer the reader to the previous section, and briefly highlight how these limitations apply here. Regarding normative grounding, instance-centric approaches heavily rely on the definition of individual discrimination and the specified similarity metric. For example, many of these works assume that discrimination only exists when two individuals that only differ in their sensitive features receive different outcomes (e.g. [3]). Meanwhile, applications of knowledge or causal graphs for interpretations of fairness come at the cost of making strong assumptions [49], which may not always hold in practice. Moreover, Hu and Kohler-Hausmann [104] argue that the validity of constructing counterfactuals based on sensitive features like sex may be questionable altogether, because they make flawed ontological assumptions. In particular, features that such diagrams understand as "effects" of sensitive features are in fact essential to the social meaning of said attributes.

Critique: Underappreciated forms of formal fairness analysis. Only a few articles shed light on more subtle facets of formal fairness that emerge with AI-informed decision-making. For example, Balagopalan et al. [22] and Dai et al. [58] argue that disparities in the quality ("fidelity") of explanations introduce a novel kind of formal unfairness. This adds another layer to the relationship between XAI and fairness in that explanations are not only supposed to indicate unfairness but may themselves exert a form of unfairness. Moreover, Gupta et al. [93] and Karimi et al. [109] examine the fairness of recourse; that is, explanations that provide

guidance to affected parties on how to turn a negative into a positive prediction. Recourse can be formalized by the distance to the decision boundary, which can also introduce disparities between demographic groups that conventional fairness metrics fail to reflect. By accounting for fairness of recourse, XAI might contribute to a more holistic view of formal fairness that not only considers discrete points in time but also addresses future actions of affected parties. However, Slack et al. [198] warn that existing techniques to assess recourse typically rely on counterfactual explanations, which can be manipulated, and show how XAI methods are prone to concealing group disparities in the cost of achieving recourse (i.e., how much effort has to be invested in order to change an individual's outcome).

4.4 Claim 4: *"XAI enables humans to mitigate (formal) unfairness."*

Several works observing formal unfairness directly employ countermeasures to mitigate it (e.g., [140]) or propose mitigation as a next step for future work (e.g., [143]). The sequence of (*i*) detecting, (*ii*) analyzing sources, and (*iii*) mitigating unfairness aligns with the distinction between epistemic and substantial facets of fairness desiderata [122]; where (*i*) and (*ii*) correspond to the epistemic facet, and (*iii*) corresponds to the substantial facet. In some cases, the facets can coincide when XAI methods like feature importance are directly integrated into training and bias mitigation algorithms.

> "To inhibit discrimination in algorithmic systems, we propose to nullify the influence of the protected attribute on the output of the system, while preserving the influence of remaining features" [88, p. 1].

There are several studies that employ XAI methods with the goal of mitigating formal unfairness at the pre-processing, in-processing, or post-processing stage of AI systems (e.g., [8, 101, 165]). One common approach of using XAI for unfairness mitigation is to implement "interpretable" fairness constraints during model training, which has been done for rule lists (e.g., [8]), random forests (e.g., [4]), and deep neural networks (e.g., [219]). Other methods include retraining algorithms that incorporate a fairness regularization term, which prior works compute with SHAP (e.g., [101]) or constructed counterfactuals (e.g., [59]). Aiming to resolve formal unfairness arising through concept drift, others have established a monitoring system that is claimed to automatically detect formal unfairness, attribute it to a responsible feature, and mitigate it [84]. Lastly, several articles [11–13, 28] propose feature dropout algorithms as a mitigation technique once an XAI method (e.g., LIME) detects reliance on sensitive features.

Critique: Implicit assumptions when mitigating formal unfairness. Many XAI-based unfairness mitigation strategies suffer from the same shortcomings discussed in Section 4.2, because they rely on the same normative groundings and assumptions about features. In particular, we note that many of these methods rely on the use of XAI to identify sources of fairness—and then use this information in the mitigation step. For example, when relying on LIME or SHAP to reduce the feature importance of sensitive features (e.g., [12, 13, 28], redundant encoding is a salient concern[143]. Further, employing XAI to improve measures of individual fairness

should be thoroughly grounded on normative assumptions about the source of existing group disparities [31].

Critique: Risk of exacerbating unfairness. The lack of explicit normative grounding also risks *exacerbating* bias inadvertently instead of mitigating it. Recall the discussion in Section 4.2, where we note that using the information captured in sensitive features is not always undesirable from a fairness perspective [114, 152]. Group-specific differences in the predictive relationship between covariates and outcomes are known as *differential subgroup validity* [105]. This is common across domains and can emerge for different reasons. For instance, previous patterns of discrimination may lead to differential predictive relationships between standardized test scores and college success [114]. In other cases, the nature of the problem may inherently lead to group-specific relationships. For instance, the language in hate speech is different when the target group varies, and accounting for this heterogeneity can improve overall performance and fairness of hate speech detection algorithms [92]. In such cases, applying mitigation strategies that erode group-specific patterns could exacerbate bias stemming from differential subgroup validity by, for example, favoring the patterns that are predictive for the majority group.

4.5 Claim 5: *"XAI informs human judgment of fairness."*

Whereas Section 4.2 summarizes XAI methods to provide descriptive information on formal fairness, this section discusses how humans interpret this information to make (non-formal) fairness judgments. Intuitively, if a model can justify its reasoning, a human should be able to judge whether it complies with normative standards or moral intuition.

> *"Using XAI 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"* [2, p. 52142].

Stakeholders may use information generated by XAI in multiple ways. *Deployers* are often interested in using XAI to justify the decisions of their models in order to comply with legal frameworks and to foster trust and acceptance [56]. *Regulators* are expected to establish and audit regulatory requirements on transparency and fairness to steer algorithmic decisions in a socially acceptable direction (e.g., [170]). Similarly, Leslie [127] demands deployers to prove to an external auditor that their system is "ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing." Finally, XAI aims to serve *affected parties* in multiple ways. Like other stakeholders, affected parties should be able to make well-founded judgments about model fairness, and XAI has been claimed to help in this regard (e.g., [33]). However, addressing the limited access to information and lack of AI literacy, it is frequently demanded that affected parties should receive a dedicated set of information to engage in an informed discourse (e.g., [175]). These demands are reinforced by Wachter et al. [218], calling for explanations that enable the understanding of decisions, support contestability, and provide guidance on recourse.

Critique: Disputed value of XAI for auditors. At its most general, some works argue that XAI is not required for fairness audits (e.g., [203, 226]). When the central concern of regulators is distributive fairness, relying on human judgement to assess this may not only be unnecessary but may also be misleading [181]. Moreover, several caveats on XAI for fairness reporting discussed in Section 4.2 have downstream consequences for auditing; for example, the capacity of XAI methods to intentionally or unintentionally conceal the feature importance of sensitive features [121]. "Fairwashing" loopholes [6, 65] pose threats to validity and robustness of expert audits. Recent work has also alluded to similar problems in the light of model multiplicity [34, 138]; that is, situations where different predictive models are admissible. In such cases, different models may entail different explanations, potentially allowing system designers to strategically obscure fairness concerns [34]. Generally, an alignment of normative grounding is crucial to the use of XAI for auditing. The lack of specificity of normative grounding discussed in Section 4.2 risks the misuse of XAI methods in audits where the normative grounding of the regulation is misaligned with the implicit grounding of the methodology deployed.

Critique: Disputed value of XAI for affected parties. Addressing informational needs of affected parties, XAI is often seen as a valuable tool in the contexts of the "right to explanation" (e.g., [87, 98]), contestability (e.g., [221]), and recourse (e.g., [93]). These desiderata are closely related to the concept of *informational fairness*, which has been defined as providing "adequate information on and explanation of the decision-making process and its outcomes" [183], but also to the concept of "informed self-advocacy" [217]. Prior work has observed that certain types of explanations can support humans in acknowledging formal unfairness but also stresses that the context of deployment is a crucial moderator [33, 66]. Concerningly, to date, there is deficient guidance on how to design XAI for informational fairness and informed self-advocacy. Asher et al. [21], Wachter et al. [218], and Watson and Floridi [227] provide conceptual starting points for formal requirements of explanations for affected parties, which, to our knowledge, have not been adequately tested in practice. Moreover, Schoeffer et al. [183] find that in some cases, people have no concerns with opaque decisions and raise questions about the actionability of different explanations regarding recourse. Schlicker et al. [179] find no effects of XAI on informational justice measures, either. It remains an open challenge to understand what kind of situational information XAI can provide to affected parties in order for them to *be* and *feel* treated "fairly" beyond the (currently) shallow concept of the "right to explanation" [218].

Critique: XAI can mislead human judgment. Adding to the above critiques, prior research stresses that information can be communicated in *unfair* ways. Aïvodji et al. [6] argue that the lack of specificity in XAI requirements creates incentives to provide deceptive explanations. Le Merrer and Trédan [124] formally show that remote explainability (i.e., indirect access to a single local explanation) makes it impossible for individuals to detect untruthful manipulations. John-Mathews [108] proposes the concept of denunciatory power, which describes an explanation's capacity to reveal an "unfair" incident. Their experimental study shows how system providers are incentivized to select the explanation with

the lowest denunciatory power to minimize negative feedback. On the other hand, misleading explanations can even occur despite benevolent intentions [70]. Providing a game-theoretic framework, Watson and Floridi [227] conceptualize accuracy, simplicity, and relevance as the key properties of fair explanations. That is, the provided information should be trustworthy, understandable, and helpful to the explainee. They formally show that even accurate explanations can be misleading if they are incomprehensible or the information content is worthless.

4.6 Claim 6: *"XAI improves human perceptions of fairness."*

Beyond formal fairness, XAI is often touted to promote positive opinions and feelings about fairness of AI systems, which is closely connected with trust and acceptance (e.g., [158, 194]).

> *"The aim of local explanations is to strengthen the confidence and trust of users that the system is not (or will not be) conflicting with their values, i.e. that it does not violate fairness or neutrality"* [170, p. 5].

In a recent survey, Starke et al. [206] find "tentative evidence that explanations can increase perceived fairness" and note that fairness perceptions are moderated by a range of factors, including the context of deployment, political ideology, AI literacy, and self-interest. To disentangle the effect of XAI on perceived fairness, several studies build on the justice constructs of Colquitt [55], which decomposes fairness perceptions into an informational, procedural, and distributive dimension. Accordingly, prior work has suggested that explaining models (*perceived informational fairness*) enables and moderates fairness judgments about the underlying process (*perceived procedural fairness*) and its outcome (*perceived distributive fairness*) [33]. There is some evidence that XAI is effective in increasing informational fairness and trustworthiness perceptions, even over explanations provided by human decision-makers [183, 185]. However, findings on perceived procedural and distributive fairness are mostly inconclusive (e.g., [33, 179]). This might be due to the dual effect of XAI on perceived fairness described by Lee et al. [125]: XAI can contribute to more understanding and transparent treatment (which relates to informational fairness); at the same time, XAI can unveil properties of the model that might conflict with people's fairness beliefs (which relates to procedural or distributive fairness).

Critique: Societal concerns with maximizing perceived fairness. Positive fairness perceptions may, in several cases, be desirable but can emerge for questionable reasons. For example, Shulner-Tal et al. [193] find that the effect of explanations on perceived fairness is primarily dominated by outcome favorability. Moreover, negative outputs tend to be regarded as unfair, regardless of the explanation [193]. Shin [191] finds that the mere act of providing explanations positively affects source credibility, which makes humans prone to form trust based on placebic [71] or manipulative explanations [6, 121]. Similarly, it has been shown that explanations can increase participants' trust and fairness perceptions even if the scenario primes the model as unfair [16]. From a societal perspective, this is concerning because users might inappropriately rely on unfair model output, and affected parties might not recognize that they are treated unfairly. Therefore, a key desideratum of XAI in many cases may not be to foster *positive* fairness perceptions but *appropriate* (i.e., calibrated) fairness perceptions [182] instead.

4.7 Claim 7: *"XAI enables humans to implement subjective notions of fairness."*

It has been claimed that stakeholders can adjust a model towards non-formalized notions of fairness based on factors such as morale, domain-specific expertise, or other contextual factors. Throughout a series of co-design workshops, Stumpf et al. [209] especially highlight users and affected parties as key stakeholders to mitigate unfairness by incorporating feedback into the model. From the perspective of users, Schoeffer et al. [181] suggest that XAI should enable humans to calibrate their reliance behavior accordingly, but also identify a lack of empirical evidence for this. An exemplary claim in support of this idea is Chakraborty et al. [47] proposing an XAI method that visualizes the nearest neighbors of an unfairly classified data point.

> *"We generate this tabular explanation for all test data points which are unfairly treated. A domain expert can easily evaluate our explanations and take decision whether to change the prediction or not"* [47, p. 1231].

Moreover, XAI has been claimed to enable domain experts to make better trade-off decisions, for example, between fairness and accuracy (e.g., [5, 7]). Some works have also proposed to have users directly incorporate domain-specific interpretable constraints into the model (e.g., [235]). Others support the idea of actively integrating XAI-based feedback on fairness from affected parties into the design process of a model (e.g., [76, 209]). Finally, prior work has employed XAI to explore how affected parties can identify unfair use of features that developers might overlook or directly implement feedback themselves (e.g., [149, 213]).

Critique: Threat of ineffective "humans-in-the-loophole". Existing laws and regulations (e.g., the GDPR), assign an essential role to a human-in-the-loop as a safeguard for fairness and accountability [98]. In fact, prior work has argued that human points of contact are valuable for a sense of interpersonal fairness [55]. Also, human discretion may be required to make normative trade-offs [186] and to overrule intolerable outputs [221]. However, humans engaging in AI-informed decision-making should be provided with adequate tools to foster effective and responsible reliance behavior [25, 180]. Otherwise, real-world applications might be at risk of installing ineffective "humans-in-the-loophole" [32] that legitimate whatever the underlying logic of the model dictates. There is a need for effective XAI tools to capitalize on the complementary capabilities of humans and AI systems for fairness desiderata—for example, in cases where AI systems are superior in analyzing statistical patterns, and human discretion balances fairness desiderata based on the societal context. However, empirical evidence suggests that the practical benefits of existing XAI methods for the human-in-the-loop are limited [181]. Instead, prior work has shown that XAI may often lead to situations where humans under- or over-rely on AI advice [178].

Critique: XAI is not providing relevant cues to human stakeholders. XAI techniques need to be designed to offer meaningful

cues to human stakeholders, empowering them to effectively exercise their discretionary power in pursuit of specific objectives. With respect to formal distributive unfairness, recent work highlights concerns that widely used XAI methods such as LIME and SHAP may not provide such cues [181]. Instead, these techniques merely indicate whether an AI system utilizes sensitive features, which is not a reliable signal for fairness because of redundant encoding [65, 143] or differing base rates [69, 131], as discussed in Section 4.2. Emerging XAI techniques should empower human stakeholders to base their discretion on information that is both relevant and reliable for enhancing fairness objectives.

5 THREE PATTERNS OF CRITIQUE

According to our survey, there is a prevailing optimism in recent literature regarding XAI as a catalyst for promoting fairness in AI-informed decision-making. However, we argue that several of these optimistic expectations are misplaced. Specifically, many claims on alleged fairness benefits of XAI exhibit three distinct types of shortcomings.

First, despite being highly optimistic, we find that many claims on the relationship between XAI and fairness are **vague and simplistic**. In Section 4.1, we have seen that prior work has claimed XAI to be a necessary or even sufficient condition for fairness. However, such blanket claims are prone to promoting a misguided reliance on XAI for fairness. While prior work has shown that XAI might in some cases contribute to a better understanding of AI systems [129], it is questionable why explainablity would be a "precondition for ensuring their safety and fairness" [127] if we, for example, are only interested in fair outcomes. Such claims may sometimes be based on an implicit assumption that XAI can enable humans to audit or rectify AI systems. However, this gives rise to another limitation concerning a mismatch between the anticipated benefits and the actual capabilities of XAI techniques—which we will delve into shortly. Further, the idea of "ensuring" [127] or "guaranteeing" [20] fairness often fails to consider the multidimensional and conflicting nature of fairness [62, 79, 147]. A one-size-fits-all fairness notion simply does not exist [50, 116]. We hope that our work can cultivate a more nuanced language for future research on the potential capabilities of XAI for fairness.

Second, many fairness desiderata pursued with XAI methods are **lacking normative grounding**. For example, several articles treat "reliance on sensitive features" [12] as a form of unfairness without offering a normative rationale for why such reliance might be problematic. Binns [31] argues that this notion relates to a very confined form of unfairness that assumes a worldview where group disparities are solely due to personal choices. While this view might be valid in (hypothetical) societies where no structural disadvantages occur, it is crucial to critically reflect upon this fundamental assumption [73]. In fact, prior work has shown that actively *considering* sensitive features, such as gender or race, may significantly improve performance for historically marginalized groups like Black people and women and reduce algorithmic bias [114, 131, 139, 164, 196]; and this is closely connected to the presence of differential subgroup validity [105]. Finally, claims suggesting XAI as a means to improve fairness perceptions often fail to explain why positive fairness perceptions are a desirable goal in themselves—particularly

in light of prior work showing that human perceptions are easily misled [51, 121, 199]. Instead, in many cases, a more ethically justifiable goal might be to foster *appropriate* fairness perceptions, which are positive if and only if the underlying AI system is fair [182].

Third, even in cases of specifying and motivating a valid fairness desideratum, some claims are **poorly aligned with the actual capabilities of XAI**. For example, if the goal is to achieve formal distributive fairness, it is unclear how exactly XAI should promote this. In fact, prior research has shown that popular feature-based explanations like LIME and SHAP are unreliable mechanisms for enabling humans to enhance formal distributive fairness through leveraging their discretionary power [180]. It has also been shown that many post-hoc XAI methods are not suitable for auditing AI systems due to their susceptibility to manipulations [6, 14, 65, 198, 199]. For example, Dimanov et al. [65] show that adversarial model explanation attacks can reduce the feature importance of sensitive features and still exacerbate distributive unfairness. In some cases, XAI appears to be suited for reflecting upon the legitimacy of features (e.g., [14]), which, again, requires proper normative deliberations that go beyond what current XAI techniques alone can offer.

6 CONCLUSION AND OUTLOOK

We conducted a critical survey organizing and scrutinizing claims about the alleged fairness benefits of XAI. Despite many optimistic positions on XAI in the recent literature, we notice that the claimed fairness desiderata are often (*i*) vague and simplistic, (*ii*) lacking normative grounding, or (*iii*) poorly aligned with the actual capabilities of XAI. To facilitate a meaningful debate and move the field forward, our work stresses the importance of embedding XAI in a sociotechnical decision-making context that considers normative motivations and societal circumstances. Concretely, we emphasize the need to be more specific about **what** kind of XAI method is used and **which** fairness desideratum it refers to, **how** exactly it enables fairness, and **who** is the stakeholder that benefits from XAI. We hope that future work will build on our survey to address and overcome the fairness-related limitations of XAI.

By structuring the debate and building on the differentiation of fairness dimensions from prior literature, we hope to inspire a more nuanced and precise language for future research. A distinction between formal metrics and human perceptions helps to keep in mind that formal fairness criteria are often only meaningful when put into context and related to a moral principle. This is especially relevant for XAI-informed fairness reports, where the use of sensitive features can indicate discrimination but can just as well be normatively justified. The distinction between epistemic and substantial facets of XAI [122] clarifies the role XAI is meant to play: are XAI methods used to *explore* unfairness—or do we employ them to directly *alter* the model's fairness properties? Finally, the distinction between informational, procedural, and distributive fairness [55] allows a more nuanced look at specific aspects of an entire decision-making process. For example, informational fairness can provide an interesting perspective on XAI with a focus on *fair explanations* as opposed to fair decisions. Ultimately, we encourage future work to center around our raised questions of **what**, **which**, **how**, and **who** in order to map opportunities of XAI for different human stakeholders along the full lifecycle of AI systems.

ACKNOWLEDGMENTS

REFERENCES

[1] Behnoush Abdollahi and Olfa Nasraoui. 2018. Transparency in fair machine learning: The case of explainable recommender systems. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent* (2018), 21–35.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM.

[4] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1418–1426.

[5] 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.

[6] 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.

[7] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2021. Characterizing the risk of fairwashing. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc, 14822–14834.

[8] Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2021. FairCORELS, an open-source library for learning fair rule lists. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4665–4669.

[9] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483* (2021).

[10] Yasmeen Alufaisan, Murat Kantarcioglu, and Yan Zhou. 2021. Robust transparency against model inversion attacks. *IEEE Transactions on Dependable and Secure Computing* 18, 5 (2021), 2061–2073.

[11] Guilherme Alves, Maxime Amblard, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. 2021. Reducing unintended bias of ML models on tabular and textual data. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[12] Guilherme Alves, Vaishnavi Bhargava, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. 2020. FixOut: An ensemble approach to fairer models. https://hal.archives-ouvertes.fr/hal-03033181/

[13] Guilherme Alves, Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. 2021. Making ML models fairer through explanations: The case of LimeOut. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*. Springer, 3–18.

[14] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. *International Conference on Machine Learning* (2020), 314–323.

[15] Christopher J Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2022. Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. *Information Fusion* 77 (2022), 261–295.

[16] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579.

[17] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[18] 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*. 1–13.

[19] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards substantive conceptions of algorithmic fairness: Normative guidance from equal opportunity doctrines. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–10.

[20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel

[20] Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[21] 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.

[22] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1194–1206.

[23] 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)*.

[24] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

[25] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology* 35, 1 (2022), 12.

[26] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020).

[27] Janelle Berscheid and Francois Roewer-Despres. 2019. Beyond transparency: A proposed framework for accountability in decision-making AI systems. *AI Matters* 5, 2 (2019), 13–22.

[28] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. 2020. LimeOut: An ensemble approach to improve process fairness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 475–491.

[29] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 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*. 401–413.

[30] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.

[31] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[32] Reuben Binns. 2022. AI and the human in the loophole. RPA Humane AI UvA [Seminar presentation].

[33] 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 Conference on Human Factors in Computing Systems*. 1–14.

[34] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 850–863.

[35] 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*. 111–121.

[36] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[37] Pol Borrellas and Irene Unceta. 2021. The challenges of machine learning and their economic implications. *Entropy* 23, 3 (2021), 1–23.

[38] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 1–11.

[39] Roberta Calegari, Giovanni Ciatto, Enrico Denti, and Andrea Omicini. 2020. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information* 11, 3 (2020), 167.

[40] Roberta Calegari, Giovanni Ciatto, and Andrea Omicini. 2020. On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale* 14, 1 (2020), 7–32.

[41] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[42] 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.

[43] 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, 647–663.

[44] Corinne Cath. 2018. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 376, 2133 (2018), 1–8.

[45] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *Comput. Surveys* (2020).

[46] Juliana Cesaro and Fabio Gagliardi Cozman. 2019. Measuring unfairness through game-theoretic interpretability. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 253–264.

[47] 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.

[48] Gary K. Y. Chan. 2022. AI employment decision-making: Integrating the equal opportunity merit principle and explainable AI. *AI & SOCIETY* (2022), 1–12.

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

[50] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[51] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark patterns of explainability, transparency, and user control for intelligent systems. In *IUI Workshops*, Vol. 2327.

[52] Ylona Chun Tie, Melanie Birks, and Karen Francis. 2019. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine* 7 (2019), 2050312118822927.

[53] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Automated data slicing for model validation: A big data-AI integration approach. *IEEE Transactions on Knowledge and Data Engineering* 32, 12 (2019), 2284–2296.

[54] Nathan Colaner. 2022. Is explainable artificial intelligence intrinsically valuable? *AI & SOCIETY* 37, 1 (2022), 231–238.

[55] Jason 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.

[56] Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. 2021. A general model for fair and explainable recommendation in the loan domain. In *Joint Workshop Proceedings of the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021)*.

[57] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. 2022. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570* (2022).

[58] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. 2022. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 203–214.

[59] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. 2022. Alexa, in you, I trust! Fairness and interpretability issues in e-commerce search through smart speakers. In *Proceedings of the ACM Web Conference 2022*. 3695–3705.

[60] Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. 2022. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 915–924.

[61] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy*. IEEE, 598–617.

[62] 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.

[63] Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making. *AI & SOCIETY* 35, 4 (2020), 917–926.

[64] Joachim de Greeff, Maaike HT de Boer, Fieke HJ Hillerström, Freek Bomhof, Wiard Jorritsma, and Mark A Neerincx. 2021. The FATE system: Fair, transparent and explainable decision making. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 266–267.

[65] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*. IOS Press, 2473–2480.

[66] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE 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.

[67] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[68] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* 36, 4 (2020), 25–34.

[69] 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*. 214–226.

[70] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[71] 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*. 1–6.

[72] 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*. 871–882.

[73] Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

[74] Juliana Jansen Ferreira and Mateus de Souza Monteiro. 2020. Evidence-based explanation to promote fairness in AI systems. *arXiv preprint arXiv:2003.01525* (2020).

[75] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (2018), 689–707.

[76] Luciano Floridi, Josh Cowls, Thomas C King, and Mariarosaria Taddeo. 2021. How to design AI for social good: Seven essential factors. *Ethics, Governance, and Policies in Artificial Intelligence* (2021), 125–151.

[77] Danilo Franco, Nicolò Navarin, Michele Donini, Davide Anguita, and Luca Oneto. 2022. Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing* 470 (2022), 318–334.

[78] Ulrik Franke. 2022. First- and second-level bias in automated decision-making. *Philosophy & Technology* 35, 2 (2022), 21.

[79] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.

[80] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*. 577–590.

[81] Erick Galinkin. 2022. Towards a responsible AI development lifecycle: Lessons from information security. *arXiv preprint arXiv:2203.02958* (2022).

[82] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable fairness in recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 681–691.

[83] 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*. 307–316.

[84] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2023. "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 138–148.

[85] 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), 137.

[86] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89.

[87] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[88] Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1905–1916.

[89] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[90] Stephan Grimmelikhuijsen. 2023. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review* 83, 2 (2023), 241–262.

[91] Jarek Gryz and Nima Shahbazi. 2020. Futility of a right to explanation. In *EDBT/ICDT Workshops*. 72–75.

[92] 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*. 3689–3700.

[93] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166* (2019).

[94] Philipp Hacker and Jan-Hendrik Passoth. 2020. Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. In *XXAI: Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 343–373.

[95] Neal R. Haddaway, Matthew J. Grainger, and Charles T. Gray. 2022. Citation-chaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods* 13, 4 (2022), 533–545.

[96] Patrick Hall and Navdeep Gill. 2017. Debugging the black-box COMPAS risk assessment instrument to diagnose and remediate bias. https://openreview.net/pdf?id=r1iWHVJ7Z

[97] Patrick Hall, Navdeep Gill, and Nicholas Schmidt. 2019. Proposed guidelines for the responsible use of explainable machine learning. *arXiv preprint arXiv:1906.03533* (2019).

[98] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 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 (2022), 72–85.

[99] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. 2021. Amazon SageMaker Clarify: Machine learning bias detection and explainability in the cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2974–2983.

[100] Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414* (2017).

[101] James M Hickey, Pietro G Di Stefano, and Vlasios Vasileiou. 2021. Fairness by explicability and adversarial SHAP learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer, 174–190.

[102] Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2019. TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.

[103] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. GAMUT: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

[105] 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.

[106] Alexey Ignatiev, Martin C Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva. 2020. Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming*. Springer, 846–867.

[107] Aditya Jain, Manish Ravula, and Joydeep Ghosh. 2020. Biased models have biased explanations. *arXiv preprint arXiv:2012.10986* (2020).

[108] 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.

[109] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *Comput. Surveys* (2022), 1–26.

[110] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 228–236.

[111] Thomas Kehrenberg, Myles Bartlett, Oliver Thomas, and Novi Quadrianto. 2020. Null-sampling for interpretable and fair representations. In *European Conference on Computer Vision*. Springer, 565–580.

[112] Barbara Kitchenham and Stuart M. Charters. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report.

[113] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.

[114] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 108. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 22–27.

[115] Jon Kleinberg and Sendhil Mullainathan. 2019. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 807–808.

[116] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

[117] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165, 633 (2017), 633–705.

[118] Udo Kuckartz and Stefan Rädiker. 2019. *Analyzing qualitative data with MAXQDA: Text, audio, and video*. Springer.

[119] Catherine Kung and Renzhe Yu. 2020. Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*. 413–416.

[120] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[121] 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*. 79–85.

[122] 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), 103473.

[123] Connor Lawless, Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. 2023. Interpretable and fair Boolean rule sets via column generation. *Journal of Machine Learning Research* 24, 229 (2023), 1–50.

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

[125] 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. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[126] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology* 31, 4 (2018), 611–627.

[127] David Leslie. 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684* (2019).

[128] Gerald S Leventhal. 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*. Springer, 27–55.

[129] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

[130] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.

[131] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems* 31 (2018).

[132] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[133] 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.

[134] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[135] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[136] Jocelyn Maclure. 2021. AI, explainability and public reason: The argument from the limitations of the human mind. *Minds and Machines* 31, 3 (2021), 421–438.

[137] Marta Marchiori Manerba and Riccardo Guidotti. 2021. FairShades: Fairness auditing via explainability in abusive language detection systems. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 34–43.

[138] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.

[139] Sandra G Mayson. 2018. Bias in, bias out. *The Yale Law Journal* 128 (2018), 2218–2300.

[140] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports* 12, 1 (2022), 7166.

[141] Loizos Michael. 2019. Machine coaching. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*. 80–86.

[142] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. 2021. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law* 29, 2 (2021), 111–147.

[143] Alexey Miroshnikov, Konstandinos Kotsiopoulos, Ryan Franks, and Arjun Ravi Kannan. 2022. Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning* (2022), 1–51.

[144] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big*

Data & Society 3, 2 (2016), 1–21.

[145] Lily Morse, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. 2021. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics* (2021), 1–13.

[146] Erica Mourão, João Felipe Pimentel, Leonardo Murta, Marcos Kalinowski, Emilia Mendes, and Claes Wohlin. 2020. On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and Software Technology* 123 (2020), 106294.

[147] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[148] Ece Çiğdem Mutlu, Niloofar Yousefi, and Ozlem Ozmen Garibay. 2022. Contrastive counterfactual fairness in algorithmic decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 499–507.

[149] 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 (TiiS)* 12, 3 (2022), 1–30.

[150] Rym Nassih and Abdelaziz Berrado. 2020. State of the art of fairness, interpretability and explainability in machine learning: Case of PRIM. In *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*. 1–5.

[151] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[152] Julian Nyarko, Sharad Goel, and Roseanna Sommers. 2021. Breaking taboos in fair machine learning: An experimental study. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[153] Alfonso Ortega, Julian Fierrez, Aythami Morales, Zilong Wang, Marina de la Cruz, César Luis Alonso, and Tony Ribeiro. 2021. Symbolic AI for XAI: Evaluating LFIT inductive programming for explaining biases in machine learning. *Computers* 10, 11 (2021), 154.

[154] Deepak Padmanabhan, V Sanil, and Joemon Jose. 2020. On fairness and interpretability. In *IJCAI 2020 AI for Social Good Workshop*.

[155] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372 (2021).

[156] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1287–1297.

[157] 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.

[158] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*.

[159] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–33.

[160] Álvaro Parafita and Jordi Vitrià. 2021. Deep causal graphs for causal inference, black-box explainability and fairness. In *Artificial Intelligence Research and Development: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*. 415–424.

[161] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[162] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[163] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. 2022. FAIR: Fair adversarial instance re-weighting. *Neurocomputing* 476 (2022), 14–37.

[164] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* 4, 7 (2020), 736–745.

[165] 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*. 247–261.

[166] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8227–8236.

[167] Bilal Qureshi, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. 2020. Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems* 54, 2 (2020), 425–437.

[168] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[169] Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 243–250.

[170] 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, 19–36.

[171] 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*.

[172] Jake Robertson, Catherine Stinson, and Ting Hu. 2022. A bio-inspired framework for machine bias interpretation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 588–598.

[173] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems* 33 (2019), 673–705.

[174] 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.

[175] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[176] 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.

[177] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity* 52 (2018), 1893–1907.

[178] 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.

[179] Nadine Schlicker, Markus Langer, Sonja K. Ötting, Kevin Baum, Cornelius J. König, and Dieter Wallach. 2021. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior* 122 (2021), 1–16.

[180] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On explanations, fairness, and appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2209.11812* (2022).

[181] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On the relationship between explanations, fairness perceptions, and decisions. *ACM CHI 2022 Workshop on Human-Centered XAI (HCXAI)* (2022).

[182] 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*. 153–157.

[183] 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 *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[184] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 2021. A study on fairness and trust perceptions in automated decision making. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[185] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 2022. Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*.

[186] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham Law Review* 87 (2018), 1085–1139.

[187] William Seymour. 2018. Detecting bias: Does an algorithm have to be transparent in order to be fair? *BIAS 2018* (2018).

[188] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 166–172.

[189] Donghee Shin. 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 (2020), 541–565.

[190] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021).

[191] Donghee Shin. 2021. Why does explainability matter in news analytic systems? Proposing explainable analytic journalism. *Journalism Studies* 22, 8 (2021), 1047–1065.

[192] Donghee Shin, Joon Soo Lim, Norita Ahmad, and Mohammed Ibahrine. 2022. Understanding user sensemaking in fairness and transparency in algorithms: Algorithmic sensemaking in over-the-top platform. *AI & SOCIETY* (2022), 1–14.

[193] 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.

[194] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 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 (2023), 1455–1482.

[195] Michael Siering. 2022. Explainability and fairness of RegTech for regulatory enforcement: Automated monitoring of consumer complaints. *Decision Support Systems* 158 (2022), 1–12.

[196] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. 2016. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior* 40, 5 (2016), 580.

[197] Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and Fair-MAML: Learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 200–209.

[198] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems* 34 (2021), 62–75.

[199] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 180–186.

[200] Kacper Sokol and Peter Flach. 2019. Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In *2019 AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019.*

[201] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 56–67.

[202] Timo Speith. 2022. How to evaluate explainability? A case for three criteria. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW).* IEEE, 92–97.

[203] Aaron Springer and Steve Whittaker. 2019. Making transparency clear. In *Joint Proceedings of the ACM IUI 2019 Workshops.*

[204] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2459–2468.

[205] 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), 20539517221115189.

[206] 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* 3, 1 (2022), 303–314.

[207] Cor Steging, Silja Renooij, and Bart Verheij. 2021. Discovering the rationale of decisions: Towards a method for aligning learning and reasoning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law.* 235–239.

[208] Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen. 2020. Explainability and fairness in machine learning: Improve fair end-to-end lending for Kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI).* IEEE, 1241–1248.

[209] 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.*

[210] Gero Szepannek and Karsten Lübke. 2021. Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence* 4 (2021), 681915.

[211] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law.* 83–92.

[212] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P).* IEEE, 401–416.

[213] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[214] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[215] Lingraj S Vannur, Balaji Ganesan, Lokesh Nagalapatti, Hima Patel, and MN Tippeswamy. 2021. Data augmentation for fairness in personal knowledge base population. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25.* Springer, 143–152.

[216] Carla Piazzon Vieira and Luciano Antonio Digiampietri. 2022. Machine learning post-hoc interpretability: A systematic mapping study. In *Proceedings of the XVIII Brazilian Symposium on Information Systems.* 1–8.

[217] Kate Vredenburgh. 2022. The right to explanation. *Journal of Political Philosophy* 30, 2 (2022), 209–229.

[218] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[219] B. Wagner and A. S. d'Avila Garcez. 2021. Neural-symbolic integration for fairness in AI. In *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering.*

[220] Robyn Repko Waller and Russell L Waller. 2022. Assembled bias: Beyond transparent algorithmic bias. *Minds and Machines* 32, 3 (2022), 533–562.

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

[222] Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. 2023. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology* 39, 2 (2023), 519–581.

[223] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Shapley Q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 34. 7285–7292.

[224] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

[225] Ziming Wang, Changwu Huang, and Xin Yao. 2024. Procedural fairness in machine learning. *arXiv preprint arXiv:2404.01877* (2024).

[226] Richard Warner and Robert H Sloan. 2021. Making artificial intelligence transparent: Fairness and the problem of proxy variables. *Criminal Justice Ethics* 40, 1 (2021), 23–39.

[227] 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.* Springer, 185–219.

[228] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering.* ACM, 1–10.

[229] Claes Wohlin, Marcos Kalinowski, Katia Romero Felizardo, and Emilia Mendes. 2022. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology* 147 (2022), 106908.

[230] Joost F. Wolfswinkel, Elfi Furtmueller, and Celeste P. M. Wilderom. 2013. Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems* 22, 1 (2013), 1–11.

[231] Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, and Finale Doshi-Velez. 2020. Incorporating interpretable output constraints in Bayesian neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 12721–12731.

[232] Carlos Vladimiro González Zelaya. 2019. Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE).* IEEE, 2086–2090.

[233] Hantian Zhang, Nima Shahbazi, Xu Chu, and Abolfazl Asudeh. 2021. FairRover: Explorative model building for fair and responsible machine learning. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning.* 1–10.

[234] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making – The causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 32.

[235] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 295–305.

[236] Haibin Zheng, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. 2022. NeuronFair: Interpretable white-box fairness testing through biased neuron identification. In *Proceedings of the 44th International Conference on Software Engineering*. 1519–1531.

[237] Jianlong Zhou, Fang Chen, and Andreas Holzinger. 2020. Towards explainability for AI fairness. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 375–386.

[238] Julian Zucker and Myraeka d'Leeuwen. 2020. Arbiter: A domain-specific language for ethical machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 421–425.

A METHODOLOGIES OF SURVEYED LITERATURE

Consistent with Langer et al. [122], the literature identified from our systematic review is highly diverse with respect to methodologies, pursued desiderata, and addressed stakeholders. To provide an overview of the examined set of articles, we briefly summarize the methodologies used in these articles. The key goal of this is not a perfectly distinct categorization but rather an emphasis on the types of evidence for the respective claims. Table 2 breaks down the methodologies used in the 175 reviewed articles and provides prominent examples to clarify the categories. Note that the counts add up to more than 175 due to some articles using more than one method. For example, Ahn and Lin [5] propose a design framework, instantiate it on real-life data, and additionally conduct user studies to demonstrate its use for practitioners.

Table 2: Methodologies used in the 175 reviewed articles.

Methodology	Count	Exemplary articles
Conceptual	**76**	
Framework	35	[75, 115, 122]
Argumentation	24	[117, 132, 174]
Literature review	20	[20, 67, 126]
ML evaluation	**84**	
XAI/fairness method	63	[61, 88, 234]
Case study	12	[85, 119, 142]
Applied framework	9	[5, 99, 188]
Human subject studies	**29**	
Quantitative study	23	[33, 66, 108]
Qualitative study	14	[66, 125, 184]

Conceptual contributions comprise literature reviews and argumentation (such as position papers) building on prior work and reasoning. Elaborate recommendations for design, evaluation, or regulation, as well as conceptual or formal models also fall into this category, labeled as frameworks. *ML evaluation* work comprises all studies that empirically evaluate a machine learning (ML) method or framework on real-world datasets. The most prevalent type of research is the empirical evaluation of an XAI and/or fairness method. This also includes work that scrutinizes existing XAI methods by performing adversarial attacks. Case studies applying existing methods in a specific domain or context are also included in this category. Further, if a framework is empirically evaluated on data, it additionally appears in this category as applied framework. Finally, *human subject studies* involve empirical examination of human perceptions, needs, or feedback. While quantitative methods

mostly test statistical significance of hypotheses on fairness perceptions, qualitative studies typically explore reasoning and opinions of various stakeholders.

B REFERENCES FOR ARCHETYPAL CLAIMS

Table 3 provides a comprehensive overview of references organized along archetypal claims, distinguishing between supportive and critical perspectives towards these claims.

Table 3: Overview of references for all seven archetypal claims.

Archetypal claim	Stance	References
"XAI helps achieve (a generic notion of) fairness."	**Supportive**	[1, 2, 10, 20, 39, 40, 44, 54, 74, 81, 85, 127, 150, 176, 200, 207, 216]
	Critical	[122, 144]
"XAI enables humans to report on (formal) fairness."	**Supportive**	[1, 3, 9, 12, 13, 15, 20, 26, 27, 35, 37, 43, 46, 59, 61, 64, 67, 68, 72, 77, 78, 80, 85, 86, 88, 94, 96, 99, 102, 103, 106, 107, 122, 132, 137, 140–144, 153, 157, 158, 160, 163, 166, 167, 169, 170, 173, 174, 188, 200, 208, 210–212, 215, 216, 222, 233]
	Critical	[7, 9, 23, 26, 65, 88, 96, 117, 140, 187]
"XAI enables humans to analyze sources of (formal) fairness."	**Supportive**	[1, 3, 5, 9, 14, 20, 22, 26, 35, 47, 53, 58, 61, 68, 72, 80, 82–85, 88, 127, 137, 142, 143, 148, 153, 156, 157, 165–167, 169, 172, 195, 199, 211, 212, 215, 232–234, 236, 237]
	Critical	[9, 226].
"XAI enables humans to mitigate (formal) unfairness."	**Supportive**	[4, 7, 11–13, 28, 29, 59, 64, 72, 78, 82, 83, 88, 93, 96, 97, 101, 102, 107, 111, 122, 123, 130, 140, 143, 148, 156, 160, 165, 175, 195, 211, 219, 223, 231, 234–236]
	Critical	[78, 220]
"XAI informs human judgement of fairness."	**Supportive**	[2, 16, 18, 21, 27, 33, 44, 48, 56, 63, 66, 67, 75, 78, 80, 85, 90, 93, 94, 97, 98, 101, 102, 108, 109, 113, 124, 125, 127, 133, 133, 136, 149, 154, 159, 169, 175, 183, 187, 188, 190, 193, 201, 203, 214, 218, 226, 227, 238]
	Critical	[6, 14, 16, 22, 26, 33, 47, 48, 57, 63, 65, 81, 85, 86, 90, 91, 100, 108, 110, 117, 124, 126, 133, 136, 144, 174, 181, 182, 186, 187, 191, 194, 198, 199, 203, 218, 221, 226, 227]
"XAI improves human perceptions of fairness."	**Supportive**	[16, 18, 33, 66, 108, 125, 158, 161, 162, 170, 182–185, 190–194, 205, 214, 224]
	Critical	[33, 125, 168, 179, 183, 192, 205]
"XAI enables humans to implement subjective notions of fairness."	**Supportive**	[4, 5, 7, 47, 48, 64, 66, 76, 119, 136, 149, 157, 162, 197, 209, 213, 219, 231, 233, 235, 238]
	Critical	[149, 181]