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
deartega@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 sub-debates, 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:

\[
(xai \text{ OR explanation OR understandab}^* \text{ OR intelligib}^* \text{ OR comprehensib}^* \text{ OR interpretab}^* \text{ OR explainab}^* \text{ OR transparent}^*) \text{ AND fair}^* \text{ AND (ai OR artificial intelligence OR machine learning)}
\]

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. We allow for snowballing to include publications outside of Scopus using the Lens.org database. 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 [135], 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
Identification of studies via databases and registers

Records identified from:
- Scopus (n=909)
- arXiv.org (n=94)

Records removed before screening:
- Non-paper (n=63)
- Prior to 2016 (n=53)
- Non-English (n=12)
- Duplicates (n=50)

Identification via other methods

Records identified from:
- Backward snowballing (n>4000)
- Forward snowballing (n>10,000)

Iteration 1: 170 articles
122 +28 +20 +0 +5

BS1 BS2 FS1 FS2

Iteration 2: 175 articles

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
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.

<table>
<thead>
<tr>
<th>Archetypal claim</th>
<th>Stance</th>
<th>Exemplary quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>“XAI helps achieve (a generic notion of) fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“[A] perfectly auditable algorithmic decision [...] can nevertheless cause unfair and transformative effects” [144, pp. 14–15].</td>
</tr>
<tr>
<td>“XAI enables humans to report on (formal) fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“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].</td>
</tr>
<tr>
<td>“XAI enables humans to analyze sources of (formal) fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“[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].</td>
</tr>
<tr>
<td>“XAI enables humans to mitigate (formal) unfairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“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].</td>
</tr>
<tr>
<td>“XAI informs human judgement of fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“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].</td>
</tr>
<tr>
<td>“XAI improves human perceptions of fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“When explanations helped respondents understand biased distributions, perceived fairness decreased” [205, p. 7].</td>
</tr>
<tr>
<td>“XAI enables humans to implement subjective notions of fairness.”</td>
<td>Supportive</td>
<td>“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].</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>“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].</td>
</tr>
</tbody>
</table>

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 now 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 may only be provided by model-specific explanations [41], such as intrinsically interpretable rule lists (e.g., [8]).

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, Balogopalan 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., [149]) 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]. “Fair-washing” 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 placebo [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 explainability 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

This research was supported in part by NIH grant R01NS124642, by a Google Award for Inclusion Research, and by Good Systems, a UT Austin Grand Challenge to develop responsible AI technologies. The authors are grateful for the ongoing support from Gerhard Satzer and the KSRI team at the Karlsruhe Institute of Technology.

REFERENCES

[40] Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with...
Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A
Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of inter-
Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey
Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You
Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial intelligence, trans-
Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Eui-
Michael Chromik, Malin Eiband, Sarah Theresa Volkel, and Daniel Buschek. 2019.

[50] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining black-

[59] Sina Fazelpour and Zachary CLipton. 2020. Algorithmic fairness from a non-
[57] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey

[37] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubrama-

[28] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubrama-
[27] Josephine Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Eui-

[23] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubrama-

[21] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian
[20] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian

[18] Sina Fazelpour and Zachary CLipton. 2020. Algorithmic fairness from a non-
[17] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian

[14] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian

[12] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian

[8] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian


[1] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian
A Critical Survey on Fairness Benefits of Explainable AI


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.

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

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.

<table>
<thead>
<tr>
<th>Archetypal claim</th>
<th>Stance</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>“XAI helps achieve (a generic notion of) fairness.”</td>
<td>Supportive</td>
<td>[1, 2, 10, 20, 39, 40, 44, 54, 74, 81, 85, 127, 150, 176, 200, 207, 216]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[122, 144]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[7, 9, 23, 26, 65, 88, 96, 117, 140, 187]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[9, 226]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[78, 220]</td>
</tr>
<tr>
<td>“XAI improves human perceptions of fairness.”</td>
<td>Supportive</td>
<td>[16, 18, 33, 66, 108, 125, 158, 161, 162, 170, 182–185, 190–194, 205, 214, 224]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[33, 125, 168, 179, 183, 192, 205]</td>
</tr>
<tr>
<td>“XAI enables humans to implement subjective notions of fairness.”</td>
<td>Supportive</td>
<td>[4, 5, 7, 47, 48, 64, 66, 76, 119, 136, 149, 157, 162, 197, 209, 213, 219, 231, 233, 235, 238]</td>
</tr>
<tr>
<td></td>
<td>Critical</td>
<td>[149, 181]</td>
</tr>
</tbody>
</table>