A Critical Survey on Fairness Benefits of XAI

Luca Deck
University of Bayreuth
luca.deck@uni-bayreuth.de

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

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

Niklas Kühl
University of Bayreuth
kuehl@uni-bayreuth.de

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 papers 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. While the literature often suggests XAI to be an enabler for several fairness desiderata, we notice a divide between these desiderata and the capabilities of XAI. We encourage to conceive XAI as one of many tools to approach the multidimensional, sociotechnical challenge of algorithmic fairness and to be more specific about *how* exactly *what* kind of XAI method enables *whom* to address *which* fairness desideratum.

1 Introduction

The integration of AI into decision-making processes has raised concerns about reinforcing societal inequalities [17, 126]. Moreover, much progress in AI comes at the cost of increased complexity and opacity, which impedes human understanding [34]. Explainable AI (XAI) is commonly conceived as a remedy to both of these emerging challenges [23]. However, the implicit link between XAI and fairness has been challenged due to inconclusive evidence and a lack of consistent terminology [102, 21]. Our critical survey explores the complex relationship between XAI and algorithmic fairness by reviewing 175 papers 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 potential and limitations of XAI for fairness. While the literature suggests broad applicability of XAI for fairness, we notice a misalignment between fairness desiderata and the actual capabilities of XAI. Many claims in the literature remain vague about how exactly XAI will contribute to fairness and disregard technical limitations, conflicts of interest between stakeholders, and normative grounding.

2 Background & Methodology

Background Both XAI and fairness are multifaceted concepts with different XAI desiderata (e.g., among stakeholders [102]) and different fairness dimensions (e.g., formal vs. perceived [174]). In this work, we understand XAI as any mechanism that "produces details or reasons

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

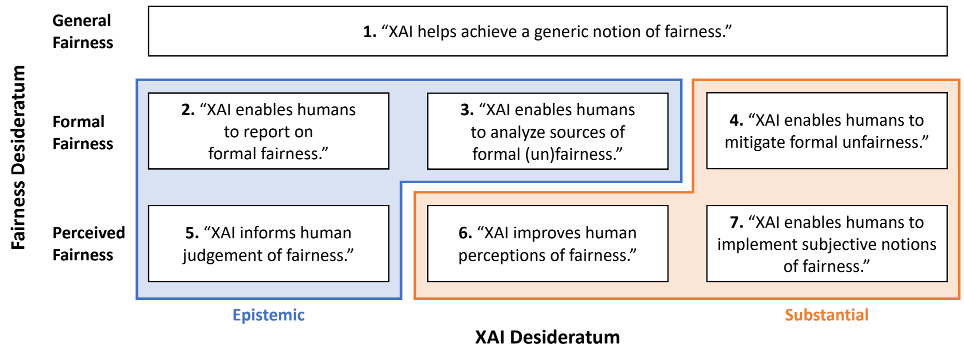

Figure 1: The seven identified archetypal claims on XAI desiderata organized along fairness dimensions.

to make [the] functioning [of an AI system] clear or easy to understand" [23]. For fairness, we distinguish between formalized notions of fairness (*formal fairness*) and human perceptions (*perceived fairness*), similar to Starke et al. [174]. Formal fairness criteria, such as statistical parity, are captured in mathematical and statistical frameworks [22] that can but do not necessarily correspond to human fairness perceptions [55, 127]. Because perceived fairness is highly context-sensitive [174] and related to complex moral deliberations [30], it requires fundamentally different measurements; for instance, based on psychological constructs [46]. Regarding XAI desiderata, we delineate *substantial* from *epistemic* facets [102]. The substantial facet refers to actual model properties (e.g., a model's reliance on sensitive attributes) whereas the epistemic facet refers to the capability of humans to observe the substantial facet (e.g., XAI providing insights into a model's reliance on sensitive attributes). This distinction is helpful to understand the multifaceted role of XAI across many application contexts. For instance, an epistemic usage of XAI is to *inform* about a given fairness desideratum, whereas a substantial usage of XAI aims at directly or indirectly *affecting* (un)fairness.

Methodology Similar to Blodgett et al. [32], we systematically identified and scrutinized claims of recent publications. We first conducted a structured literature review guided by Kitchenham and Charters [95] to identify entrenched claims on the alleged capabilities of XAI for fairness. This process yielded 175 papers, listed in Section 5.1. We supplemented our deductive literature review with inductive coding [197] at the level of individual claims. A rigorous qualitative analysis of these claims using a grounded theory approach [43] yielded seven archetypal claims, summarized in Figure 1.

3 Critical Survey

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 Section 5.3 for a complete overview of archetypal claims including all references.

3.1 Claim 1: *"XAI helps achieve a generic notion of fairness."*

Intuition 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 identified two tendencies. The first suggests XAI as a *necessary* condition for fairness:

> *"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"* [107, p. 40].

XAI is sometimes even treated as *sufficient* for achieving fairness:

> *"[F]rom a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models"* [23, p. 9].

Critique: Futility and danger of vague claims on generic fairness notions Fairness is a multidimensional concept [46, 174], so it is easy to come up with counterexamples for such strong, overgeneralized claims. For instance, distributions of classification rates can be used to show that models conform with formal fairness criteria without popular XAI tools [42]. Moreover, a transparent model can be perfectly scrutable and still be deemed unfair by some stakeholders, which precludes the suggested sufficiency [120]. 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. [62]. 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 essential needs of relevant stakeholders [102]. Perpetuating these overly general claims threatens to produce unwarranted trust and reliance towards current XAI technology.

3.2 Claim 2: *"XAI enables humans to report on formal fairness."*

Intuition A large share of papers is concerned with using XAI to measure and report on formal (un)fairness, often phrased as "identifying bias" [72] or "detecting discrimination" [90]. The central intuition behind these claims is that conventional evaluation measures (such as AUROC) may only describe a model's outcome fairness for a certain test set but fail to consider the underlying mechanisms leading to this outcome [110]. XAI is hoped to fill this gap by providing insights into these mechanisms, which can then be related to formal fairness criteria [56], such as statistical parity [90], equality of opportunity [39], or counterfactual fairness [66]. Since the anti-discriminatory motivation of formal fairness criteria typically relates to sensitive attributes such as gender or race, XAI is commonly employed to examine how models make use of these sensitive attributes.

> *"[F]eature importance measures are connected both with consistency and equality of opportunity. Consequently we see that feature importance measures do quantify both group and individual fairness"* [39, p. 9].

Prominent XAI methods to examine the use of sensitive attributes are feature importance measurements using LIME or SHAP (e.g., [39, 90, 13]) or counterfactuals (e.g., [171, 160, 66]). Moreover, inherently interpretable models are employed to gauge the use of sensitive attributes (e.g., [143, 180, 116]).

Critique: Technical and normative limitations of XAI-enabled fairness reporting In a plea for intrinsically interpretable models, Rudin [147] 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 the susceptibility of such approaches to adversarial attacks on fairness reports, which produce innocuous explanations for (formally) unfair models. Major limitations that have been addressed are *fairwashing* through rationalized surrogate models [6], reliance on input perturbations [169], or exploitation of redundant encodings [54]. Even beyond these intentional manipulations, one may challenge the framing of low feature importance of sensitive attributes as procedural fairness. First, low feature importance of sensitive attributes is neither a necessary [119, 189] nor a sufficient [25, 54] criterion to satisfy formal fairness metrics due to redundant encoding (i.e., correlated proxies for sensitive attributes) or cases of "fair affirmative action" [58]. Second, the interpretation of feature importance for fairness metrics often assumes equality of preconditions between groups (see also Section 3.4). Third, popular post-hoc XAI methods like feature importance [112] and counterfactual explanations [186] are disconnected from traditional procedural fairness notions [121]. For instance, Leventhal [108] 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) demand measures beyond formal fairness reports such as value transparency [111] and appeal processes [113].

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 a model, they shape the way a model is perceived by other 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 since explanations can be manipulated to conceal the use of sensitive attributes [6, 54, 101]. 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 [102], as well as the normative deliberations that went into the development of relevant XAI techniques [111].

3.3 Claim 3: *"XAI enables humans to analyze sources of formal (un)fairness."*

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

> "We derived the Causal Explanation Formula [...], which allows one to understand how an observed disparity between the protected attribute and the outcome variable can be decomposed in terms of the causal mechanisms underlying the specific (and unknown) decision-making process" [200, p. 2044].

This extends the epistemic facet of XAI to provide deeper-level insights 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. Some works claim to identify discriminatory samples in the training data, which serve as a basis to mitigate formal unfairness (e.g., [3, 60]). Feature-centric approaches analyze how features relate to formal fairness. Addressing the issue of redundant encoding, some works explicitly account for correlations when measuring feature influence (e.g., [51, 152]). Others extend existing feature importance methods with the goal of quantifying the influence of features on formal fairness criteria (e.g., [25, 119]). Finally, Zhang and Bareinboim [200], Grabowicz et al. [74] decompose the influence of sensitive attributes into more fine-grained effects, showing that formal unfairness can emerge subtly (e.g., via indirect or induced discrimination).

Critique: Taking a broader perspective on formal fairness Addressing problems of redundant encoding and fairwashing, some approaches in this category of claims promise to provide a more "accurate" [70] and "robust" [25] picture of the usage of sensitive features than traditional feature importance measures. However, future work must ascertain the reliability of these analytic tools. Some articles also shed light on underrepresented facets of formal fairness that emerge with AI-informed decision-making. For instance, Balagopalan et al. [20], Dai et al. [49] argue that disparities in the quality ("fidelity") of explanations introduce a novel form of formal unfairness. Moreover, Gupta et al. [78], Karimi et al. [92] examine the fairness of recourse; that is, explanations that provide guidance to affected parties on how to turn a negative into a positive prediction. Ghosh et al. [69] address the problem of intersectional fairness, which arises when multiple features are to be protected simultaneously. Finally, as Warner and Sloan [193] argue, there are facets to fairness that conventional XAI methods will never be able to reveal, such as contextual and societal factors that are not directly reflected in the data. While XAI and fairness tools (e.g., [5, 84]) should extend towards underrepresented notions of formal fairness, ethical AI frameworks (e.g., [67]), should provide guidelines to account for the societal context of fairness.

3.4 *Claim 4: "XAI enables humans to mitigate formal unfairness."*

Intuition Several papers observing formal unfairness directly employ countermeasures to mitigate it. The sequence of (i) detecting and (ii) mitigating unfairness aligns with the distinction between epistemic and substantial facets of fairness desiderata [102]. Beyond that, 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"* [74, p. 1].

There are several studies that employ XAI methods with the goal of mitigating unfairness at the pre-processing, in-processing, or post-processing stage (e.g., [86, 139, 8]). One common approach of using XAI for in-processing is to implement "interpretable" fairness constraints during model training, which has been done for rule lists [8], random forests [4], and deep neural networks [187]. Post-processing methods include retraining algorithms that incorporate a fairness regularization term, which Hickey et al. [86] compute with SHAP, and Dash et al. [50] with counterfactual explanations. Lastly, several papers [11–13, 27] propose feature dropout algorithms as a mitigation technique once an XAI method (e.g., LIME) detects reliance on sensitive features.

Critique: Process-based vs. outcome-based view Practices that rely on feature importance to nullify the impact of sensitive attributes have been challenged by drawing analogies with the notion of "fairness through unawareness" [22, 21]. As discussed in Section 3.2, the epistemic capacity of popular XAI methods to describe a model's use of protected attributes is questionable [54, 101]. And even if one can *truly* blind a model with regard to sensitive attributes, the normative motivation of blinding ought to be challenged in many cases. For instance, prior work has shown that humans often *approve of* the inclusion of sensitive attributes if it benefits historically marginalized groups [127], which questions the deontological (i.e., rule-focused) principle of blinding. Although "in practice, the process- and outcome-based views often align" [127], the interpretation of feature importance for formal individual and group fairness is heavily contingent on existing group inequalities [39]. For instance, if demographic groups have differing dispositions, the fairness notion of statistical parity can only be reached if the disadvantaged group receives preferential treatment [58], which warrants the proactive use of sensitive information for "fair affirmative action." Thus, Green [75] warns that ignoring sensitive attributes in the presence of unequal group dispositions is prone to perpetuate existing inequalities.

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

Intuition Whereas Section 3.2 summarizes XAI methods to provide descriptive information on formal fairness, this section discusses how humans make sense of this information to form 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. First, *deployers* are interested in XAI to justify the decisions of their models in order to comply with legal frameworks and to foster trust and acceptance [47]. *Regulators* establish regulatory requirements on transparency and fairness to steer algorithmic decisions into a socially acceptable direction [144]. For instance, human auditors might rely on XAI to judge the compliance with such requirements [194]. Lastly, *affected parties* are addressed by XAI in multiple ways. Like other stakeholders, affected parties should also be able to make well-founded judgments about model fairness [30]. 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 [148]. These demands are reinforced by Wachter et al. [186], calling for explanations that enable the understanding of decisions, support contestability, and provide guidance on recourse.

Critique: The disputed value of XAI for auditing and right to explanation Several caveats on XAI for fairness reporting discussed in Section 3.2 have downstream consequences for auditing; for instance, the capacity of XAI to conceal biases. Also, delineating explainability from auditability, some works argue that XAI is not required to audit formal (distributive) fairness [173, 193]; for instance, if inequalities of error rates are assessed instead. Auditors should be aware of these limitations and embed the procedural insights provided by XAI in a broader sociotechnical framework to balance societal impacts of AI-informed decision-making [111]. XAI is further seen as a valuable tool in the contexts of right to explanation [83], contestability [189], and recourse [78]. However, Aïvodji et al. [6] argue that the lack of specificity in XAI requirements creates loopholes to provide deceptive explanations to affected parties. For instance, John-Mathews [91] shows how deployers are incentivized to provide explanations that minimize negative feedback. Watson and Floridi [194] show that even accurate explanations can be misleading if they are incomprehensible or irrelevant to the explainee.

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

Intuition Beyond formal fairness, XAI is often touted to promote positive opinions and feelings about fairness of AI systems [165]. While Starke et al. [175] find "tentative evidence that explanations can increase perceived fairness," it is noted 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 decompose fairness perceptions into an informational, procedural, and a distributive dimension [46]:

> *"Requiring organisations to explain the logic behind their algorithmic decision-making systems (informational justice) enables affected individuals to assess whether the logic of the system is just (procedural justice), which in turn might moderate their assessments of fairness of the decision outcomes (distributive justice)"* [30, p. 3].

Prior work has observed that certain types of explanations support humans in acknowledging formal unfairness but also find that the context of deployment is a crucial moderator [30, 55]. There is also evidence that XAI is effective in increasing perceived informational fairness and trustworthiness, even over explanations provided by human decision-makers [156, 157]. Other works' findings on perceived procedural and distributive fairness are mostly inconclusive [30, 150]. This might be due to the dual role of XAI for perceived fairness [105]: 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: The societal concern with maximizing perceived fairness Positive fairness perceptions may in several cases be desirable but can emerge for questionable reasons. For instance, Shulner-Tal et al. [166] find that the effect of explanations on perceived fairness is primarily dominated by outcome favorability. Contrarily, negative outputs are generally regarded as unfair, regardless of the explanation [166]. Shin [163] finds that the mere act of providing explanations positively affects source credibility, which makes humans prone to form trust based on placebic [59] or manipulative explanations [101, 6]. 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 [151].

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

Intuition 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. Stumpf et al. [178] highlight users and affected parties as key stakeholders to mitigate unfairness by incorporating feedback into the model.

> *"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"* [40, p. 1231].

Further, XAI might enable domain experts to make better tradeoff decisions, for instance, between fairness and accuracy [7, 5]. Some works have also proposed to have users directly incorporate domain-specific interpretable constraints into the model [201]. Moreover, some papers support the idea of actively integrating XAI-based feedback on fairness from affected parties into the design process of a model [63, 178].

Critique: The threat of uninformed "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 [83]. Arguably, human points of contact are valuable for a sense of interpersonal fairness [46]. Also, human discretion may be required to make normative tradeoffs [158] and to overrule intolerable outputs [189]. However, humans engaging in AI-informed decision-making should be provided with adequate tools to foster effective and responsible reliance behavior

[155, 24]. Otherwise, real-world applications might be at risk of installing uninformed "humans-in-the-loophole" [29] that legitimate whatever the underlying logic of the model dictates. We are in need of effective XAI tools to capitalize on the complementary capabilities of humans and AI for fairness tasks.

4 Conclusion

We conducted a critical survey organizing and scrutinizing claims about the fairness benefits of XAI. Our work provides a comprehensive mapping of support and caveats that is meant to retrace and shape a meaningful debate. We find that fairness desiderata of XAI are often misaligned with the capabilities of XAI. For instance, feature importance is not suited for certification of fairness but can rather serve as an entry point to explore statistical relationships that might point towards unfair tendencies. Further, XAI should not be misused as a vehicle to promote positive fairness perceptions (towards potentially unfair models) but rather empower stakeholders to make well-informed judgments. We highlight that the information provided by XAI is always to be interpreted in a sociotechnical decision-making context that should consider normative motivations and societal circumstances. For future work, we, therefore, call for more clarity about *how* exactly *what* kind of XAI method enables *whom* to address *which* fairness dimension.

5 Supplementary Material

5.1 Elaborations on Methodology

5.1.1 Systematic Literature Review

In order to receive 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. Generally, our search string was supposed to reflect various dimensions of both XAI and fairness and to restrict the results to AI contexts. Beyond *explainable AI* and its acronym *XAI*, we relied on [23] and incorporated the related terms *understandability, comprehensibility, interpretability, explainability, and transparency*. Additionally, we included the frequently used keyword *explanation*. Altogether, these terms should reflect the manifold nature and discordant definitions of XAI in the literature. Fairness touches upon strongly related concepts like *discrimination, justice, ethics, or bias*, which are sometimes used synonymously. Also, some articles specifically address the opposite term unfairness. However, we found that *fair* (or compound words) appeared to be the dominant wording used in all relevant papers from our exploratory review. To narrow down towards the field of AI, we included the straightforward terms *artificial intelligence* and *machine learning*. Related terms such as *algorithm* and *automated* did not improve the quality of results and were thus left out. After screening around 400 individual papers, we finally decided on the following search string (note that the asterisk as wildcard character allows us in each case to consider both adjective and noun):

(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 the two popular search strategies database query and snowballing [196], we followed proven guidelines for systematic literature reviews in the domain of software engineering [95, 195]. Scopus was the natural choice for our database search because it is provably an effective tool to generate seed sets for snowballing [122] and includes all relevant publishers for our task, except for arxiv. To account for recent, unreviewed publications, we applied our search string to the arxiv database (limiting the search to keywords due to technical limitations of the search feature). Following the documentation guidelines of [95] and the PRISMA standard [130], we aim to provide a transparent and replicable documentation of the selection process. Figure 2 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.

At the initial identification level we only considered full papers and excluded records like courses, keynotes, etc. Here, we manually inspected all abstracts and only retained papers examining dimensions of both fairness and XAI fitting into the broader scheme of our definitions. Consequently, we discarded papers having too broad or deviating notions of the XAI terms (e.g. using the term *explain* in a different context), papers using the term fair in different contexts (*fairly, FAIR principles*, etc.), and papers where fairness or XAI are not the object of research (e.g., solely mentioning the FAT principles in the introduction). Proceeding to full-text analyses, we heuristically scanned the entire paper for specific claims about XAI and fairness. Focusing on unique statements as opposed to straightforward summaries or paraphrases of previous work eliminated most of the literature reviews. Finally, we discarded papers where the direct relationship of XAI and fairness was not considered or remained too vague. For instance, [161] examines the influence of explainability and fairness on trustworthiness but does not address the interaction effect of explainability and fairness.

Starting from this seed set of 122 papers, we performed iterative backward and forward snowballing [195]. Using the citation crawling tool Citationchaser [80], which accesses the Lens.org database, we generated a comprehensive list of all unique references (backward) and citations (forward). To detect missing relevant literature in a reasonable amount of time, we sorted the results according to the number of occurrence in the citation graph and spent more time on more frequently referenced or cited papers (i.e., publications occurring once in the citation graph only received marginal attention). After repeating the procedure for the newly added papers from the first

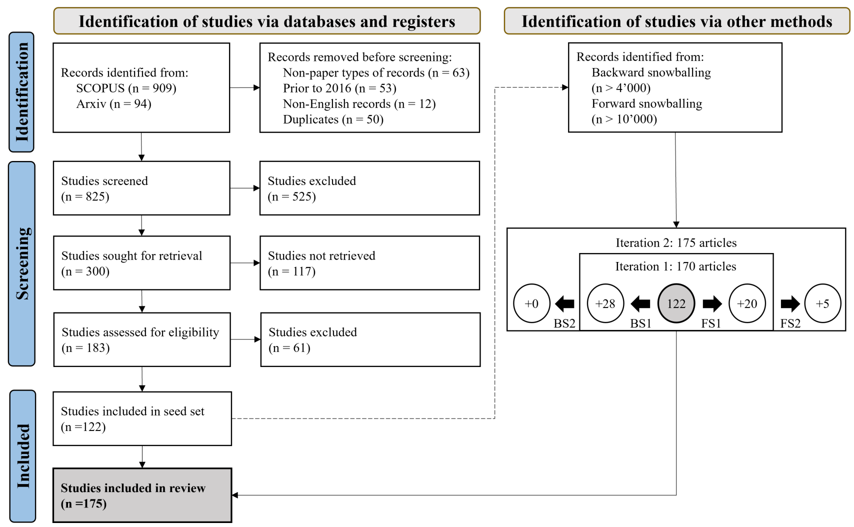

Figure 2: PRISMA flowchart describing the article selection procedure.

iteration, we found 5 additional publications. A third iteration yielded no further relevant results, which lead us to stop.

5.2 Inductive Claim Analysis

After populating our literature base, 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 [43] employing MAXQDA for claim extraction, coding, and memoing [99]. To reiterate, we built our grounded theory around the following two questions:

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

We started by skimming the full texts of our 175 selected papers to comprehend the respective methodology and key results. Parallelly, we scanned for claims with a strong focus on the most relevant and promising article sections. For example, introduction, discussion, and conclusion often provided more expressive claims than the method and results sections, which we usually only considered to retrace methodology or reasoning. However, most important claims had already been identified as part of selection criteria during the research process. 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 coder's 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 did not only consider 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 rechecking each claim.

5.3 Overview Over Archetypal Claims and Literature

Consistent with [102], the literature identified from our systematic approach is highly diverse with regards to methodologies, pursued desiderata, and addressed stakeholders. To provide an overview of the examined set of papers, we start by describing the methodology used in these

Table 1: Methodologies used in the 175 reviewed papers.

Methodology	Count	Exemplary papers
Conceptual contributions	**76**	
Framework	35	[62, 97, 102]
Argumentation	24	[98, 110, 147]
Literature review	20	[23, 56, 106]
Applied ML	**84**	
XAI/fairness method	63	[51, 74, 200]
Case study	12	[71, 100, 118]
Applied framework	9	[5, 84, 160]
Human subject studies	**29**	
Experimental study	23	[30, 55, 91]
Qualitative study	14	[55, 105, 153]

papers. The key point of this is not a perfectly distinct categorization but rather an emphasis of the types of evidence for the respective claims. Table 1 breaks down the methodologies used in the 175 papers and provides prominent examples to clarify the categories. Note that the counts add up to more than 175 due to some papers using more than one method. For example, [5] propose a design framework, instantiate it on real-life data, and additionally conduct user studies to demonstrate its use for practitioners.

Conceptual contributions comprise all studies that did not perform any primary form of empirical evaluation. Instead, this subset of papers includes literature reviews and argumentations (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. *Applied ML* work comprises all studies that empirically evaluate a method or framework on real-world datasets. By far 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 apply existing methods in a specific domain or context. 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 experimental studies quantify statistical significance of results, qualitative studies report verbatim or summarized statements of participants. From this overview, the multidimensionality of XAI and fairness becomes apparent. However, we found a methodological pivot alone to be insufficient to provide a meaningful structure for a critical discussion. Hence, we tried to group claims according to common higher-level themes.

Table 2: Overview of evidence and references for claim 1: *"XAI helps achieve a generic notion of fairness"*.

Evidence	Exemplary claims	References
Intuition	*Necessary*: "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." [107, p. 40]	[10, 107, 125, 165, 171, 176]
	Sufficient: "[F]rom a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models." [23, p. 9]	[1, 2, 23, 36, 61, 71, 185]
	Tentative: "Explainability and interpretability: these two concepts are seen as possible mechanisms to increase algorithmic fairness, transparency and accountability" [38, p. 2]	[35, 38, 45, 67, 149]
Conceptual support	"Explainability approaches may aid in this regard by providing means to track down factors that may have contributed to unfair and unethical decision-making processes and either to eliminate such factors, to mitigate them, or at least to be aware of them." [102, p. 6]	[62, 102, 120]
Conceptual caveats	"[A] perfectly auditable algorithmic decision, or one that is based on conclusive, scrutable and well-founded evidence, can nevertheless cause unfair and transformative effects, without obvious ways to trace blame among the network of contributing actors." [120, pp. 14–15]	[102, 120]

Table 3: Overview of evidence and references for claim 2: *"XAI enables humans to report on formal fairness"*.

Evidence	Exemplary claims	References
Intuition	"These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected" [72, p. 1]	[1, 23, 26, 53, 56, 57, 72, 79, 87, 88, 90, 102, 110, 116, 117, 133, 140, 144, 146, 147]
Conceptual support	"The reviewed literature showed that interpretability methods can be used for bias detection." [185, p. 6]	[33, 65, 71, 84, 171, 185]
Applied ML support	"[F]eature importance measures are connected both with consistency and equality of opportunity. Consequently we see that feature importance measures do quantify both group and individual fairness." [39, p. 9]	[3, 5, 9, 11, 13, 15, 25, 31, 37, 39, 44, 50, 51, 60, 64, 66, 70, 69, 74, 81, 84, 89, 90, 94, 115, 116, 118, 119, 128, 132, 135, 138, 141, 143, 160, 177, 179–181, 184, 190, 199]
Conceptual caveats	"The excluded or 'protected' attributes can often be implicit in other nonexcluded attributes." [98, p. 685]	[21, 74, 116, 98, 159]
Applied ML caveats	"[M]any prominent XAI tools lack features that could be critical in detecting bias." [9, p. 1]	[7, 9, 25, 54, 81, 116]

Table 4: Overview of evidence and references for claim 3: *"XAI enables humans to analyze sources of formal fairness"*.

Evidence	Exemplary claims	References
Intuition	"If your AI model is not sufficiently interpretable—if you aren't able to draw from it humanly understandable explanations of the factors that played a significant role in determining its behaviours—then you may not be able to tell how and why things go wrong in your system when they do." [107, p. 39]	[1, 3, 9, 23, 57, 107, 118, 139, 167, 180]
Conceptual support	"The investigations demonstrate that fair decision making requires extensive contextual understanding, and AI explanations help identify potential variables that are driving the unfair outcomes." [203, p. 1]	[5, 49, 71, 119, 167, 203]
Applied ML support	"We derived the Causal Explanation Formula [...], which allows one to understand how an observed disparity between the protected attribute and the outcome variable can be decomposed in terms of the causal mechanisms underlying the specific (and unknown) decision-making process." [200, p. 2044]	[3, 5, 9, 14, 20, 25, 31, 40, 44, 51, 60, 66, 68, 70, 69, 71, 73, 74, 115, 118, 119, 123, 128, 131, 139–141, 143, 145, 167, 169, 180, 181, 184, 200, 199, 202, 203]
Conceptual caveats	"[E]xplainable systems can be unfair in ways explainability will not reveal." [193, p. 31]	[193]
Applied ML caveats	"[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]	[9]

Table 5: Overview of evidence and references for claim 4: *"XAI enables humans to mitigate formal unfairness"*.

Evidence	Exemplary claims	References
Intuition	"A consequential next step to this analysis is to look for methods that mitigate unfairness in the ML methods and at the same time maintain the accuracy gains." [180, p. 7]	[28, 53, 65, 82, 87, 90, 102, 109, 116, 119, 167, 180, 187, 200]
Applied ML support	"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." [74, p. 1]	[4, 7, 11–13, 27, 50, 60, 68, 69, 74, 78, 81, 86, 103, 123, 131, 135, 139, 148, 187, 191, 198, 201, 202]
Conceptual caveats	"[W]e observe unfair recourse even when the predictions are demographically-fair." [92, p. 14]	[65, 188]

Table 6: Overview of evidence and references for claim 5: *"XAI informs human judgement of fairness"*.

Evidence	Exemplary claims	References
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]	[2, 26, 30, 38, 41, 47, 52, 55, 56, 71, 79, 83, 87, 96, 104, 111, 114, 134, 143, 148, 162, 172, 183, 186, 194, 204]
Conceptual support	"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].	[19, 30, 47, 52, 62, 65, 78, 82, 86, 91, 92, 105, 107, 111, 129, 156, 159, 166, 173, 186, 193, 204]
Applied ML support	"The explanations can be used either as justification in case the decision is challenged or as a feasible action that the individual may perform in order to improve the outcome in the future ('recourse')." [66, p. 581]	[66, 78, 160]
Human subject studies support	"Most of our participants (21/27), again regardless of expertise, mentioned that the data-centric explanations helped them judge the systems' fairness at least to some extent." [18, p. 9]	[16, 18, 30, 55, 76, 91, 105, 124, 156]
Conceptual caveats	"[T]here 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]	[6, 25, 40, 41, 48, 52, 67, 71, 72, 76, 77, 91, 85, 93, 98, 104, 106, 111, 114, 147, 151, 154, 158, 159, 163, 169, 165, 173, 186, 189, 193, 194]
Applied ML caveats	"[W]e show that it is possible to forge a fairer explanation from a truly unfair black box through a process that we coin as rationalization." [6, p. 2]	[6, 14, 20, 25, 48, 54, 169, 170]
Human subject studies caveats	"[D]epending on how and when they are deployed, explanations may or may not help individuals to evaluate the fairness of such decisions." [30, p. 10]	[6, 14, 16, 20, 30, 48, 54, 91, 169, 170]

Table 7: Overview of evidence and references for claim 6: *"XAI increases human perceptions of fairness"*.

Evidence	Exemplary claims	References
Intuition	"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" [144, p. 5]	[133, 144, 151, 163, 165, 166]
Human subject studies support	"Requiring organisations to explain the logic behind their algorithmic decision-making systems (informational justice) enables affected individuals to assess whether the logic of the system is just (procedural justice), which in turn might moderate their assessments of fairness of the decision outcomes (distributive justice)." [30, p. 3]	[16, 18, 30, 55, 91, 105, 136, 137, 153, 156, 157, 163, 162, 164–166, 183, 192]
Conceptual support	"The literature further yielded tentative evidence that explanations can increase perceived fairness." [174, p. 9]	[174]
Human subject studies caveats	"Distributive justice was not affected by the different agents and there were no effects of the types of explanations." [150, p. 13]	[30, 105, 142, 150, 156, 164]

Table 8: Overview of evidence and references for claim 7: *"XAI enables humans to implement subjective notions of fairness"*.

Evidence	Exemplary claims	References
Intuition	"There is no generally acceptable criteria[sic] for evaluating the tradeoff between fairness and utility over decision outcomes. Therefore, it is desirable to have a decision-making tool that helps incorporate the domain knowledge and human judgment to achieve fair decision making." [5, p. 9]	[5, 4, 41, 53, 55, 154]
Conceptual support	"Our work is concerned with investigating design methods for user interfaces that can help with making the fairness of AI algorithms transparent, and then help with mitigating fairness issues by incorporating user feedback back into the algorithm." [178, pp. 3–4]	[41, 63, 114, 178, 187, 199, 204]
Applied ML support	"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." [40, p. 1231]	[4, 5, 7, 40, 100, 132, 168, 198, 201]
Human subject studies support	"The perceived unfairness of specific predictors can be used to exclude predictors or identify biases in the dataset. These biases may not necessarily be apparent to those developing the AI, for example due to a lack of domain expertise, diverging social backgrounds, or personal predisposition." [182, p. 15]	[5, 124, 137, 178, 182]
Human subject studies caveats	"However, we also noted that some user input could make fairness worse. This is obviously a concern for human-in-the-loop learning as it is only as good as the input the end-user provides." [124, p. 23]	[124]

References

[1] Behnoush Abdollahi and Olfa Nasraoui. Transparency in fair machine learning: The case of explainable recommender systems. In *Human and Machine Learning*, pages 21–35. Springer, Cham, 2018.

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

[3] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 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, 2019.

[4] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 1418–1426, 2019.

[5] Yongsu Ahn and Yu-Ru Lin. FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1086–1095, 2020.

[6] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: The risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 161–170, 2019. URL http://proceedings.mlr.press/v97/aivodji19a.

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

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

[9] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. Can explainable AI explain unfairness? A framework for evaluating explainable AI. arxiv preprint, 2021. URL https://arxiv.org/abs/2106.07483.

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

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

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

[13] Guilherme Alves, Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Making ML models fairer through explanations: The case of LimeOut. In *Analysis of Images, Social Networks and Texts*, volume 12602 of *Springer eBook Collection*, pages 3–18. Springer International Publishing and Imprint Springer, 2021.

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

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

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

[17] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. In *Ethics of data and analytics*, An Auerbach Book, pages 254–264. CRC Press Taylor & Francis Group, 2022.

[18] Ariful Islam Anik and Andrea Bunt. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021.

[19] Nicholas Asher, Lucas de Lara, Soumya Paul, and Chris Russell. Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349, 2022.

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

[21] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 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)*, 2022.

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

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

[24] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):1–30, 2022.

[25] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. Explainability for fair machine learning. arxiv preprint, 2020. URL `https://arxiv.org/abs/2010.07389`.

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

[27] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. LimeOut: An ensemble approach to improve process fairness. In *ECML PKDD 2020 Workshops*, volume 1323 of *Springer eBook Collection*, pages 475–491. Springer International Publishing and Imprint Springer, 2020.

[28] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ACM Digital Library, pages 401–413. Association for Computing Machinery, 2021.

[29] Reuben Binns. AI and the human in the loophole. In *FairAI2022 Workshop*, 2022. URL `https://www.youtube.com/watch?v=XkU1-AHG1qk`.

[30] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "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*, pages 1–14. ACM, 2018.

[31] Emily Black, Samuel Yeom, and Matt Fredrikson. FlipTest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121. ACM, 2020.

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

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

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

[35] Roberta Calegari, Giovanni Ciatto, Enrico Denti, and Andrea Omicini. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information*, 11(3):1–29, 2020.

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

[37] Alessandro Castelnovo, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Cosentini. Towards fairness through time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 647–663. Springer, Cham, 2021. URL `https://link.springer.com/chapter/10.1007/978-3-030-93736-2_46`.

[38] Corinne Cath. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 376(2133):1–8, 2018.

[39] Juliana Cesaro and Fabio Gagliardi Cozman. Measuring unfairness through game-theoretic interpretability. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 253–264. Springer, Cham, 2020. URL `https://link.springer.com/chapter/10.1007/978-3-030-43823-4_22`.

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

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

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

[43] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE open medicine*, 7:2050312118822927, 2019.

[44] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Automated data slicing for model validation: A big data - AI integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2284–2296, 2020.

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

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

[47] Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. 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 RecSys 2021*, 2021.

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

[49] Jessica Dai, Sohini Upadhyay, Ulrich Aïvodji, Stephen H. Bach, and Himabindu Lakkaraju. 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*, ACM Digital Library, pages 203–214. Association for Computing Machinery, 2022.

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

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

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

[53] Joachim de Greeff, Maaike H.T. de Boer, Fieke H. J. Hillerström, Freek Bomhof, Wiard Jorritsma, and Mark Neerincx. The FATE system: Fair, transparent and explainable decision making. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2021.

[54] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *SafeAI@AAAI 2020*, 2020. URL `https://openreview.net/pdf?id=rnt7jxyjyaa`.

[55] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.

[56] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arxiv preprint, 2017. URL `https://arxiv.org/abs/1702.08608`.

[57] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2021.

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

[59] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM Digital Library, pages 1–6. Association for Computing Machinery, 2019.

[60] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. Explanation-guided fairness testing through genetic algorithm. In *Proceedings of the 44th International Conference on Software Engineering*, ACM Digital Library, pages 871–882. Association for Computing Machinery, 2022.

[61] Juliana Jansen Ferreira and Mateus de Souza Monteiro. Evidence-based explanation to promote fairness in AI systems. In *CHI2020 Fair and Responsible AI Workshop*, 2020. URL `http://arxiv.org/pdf/2003.01525v1`.

[62] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.

[63] Luciano Floridi, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, 26(3):1771–1796, 2020.

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

[65] Ulrik Franke. First- and second-level bias in automated decision-making. *Philosophy & Technology*, 35(2):1–20, 2022.

[66] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, ACM Digital Library, pages 577–590. Association for Computing Machinery, 2021.

[67] Erick Galinkin. Towards a responsible AI development lifecycle: Lessons from information security. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, 2022.

[68] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. Explainable fairness in recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22),*, 2022.

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

[70] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. "How biased is your feature?": Computing fairness influence functions with global sensitivity analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ACM Digital Library. Association for Computing Machinery, 2023.

[71] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing. *Information*, 11(3):1–32, 2020.

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

[73] Carlos Vladimiro Gonzalez Zelaya. Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th International Conference on Data Engineering*, pages 2086–2090. IEEE, 2019.

[74] Przemyslaw Grabowicz, Nicholas Perello, and Aarshee Mishra. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 2022.

[75] Ben Green. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, 35(4), 2022.

[76] Stephan Grimmelikhuijsen. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision–making. *Public Administration Review*, pages 1–22, 2022.

[77] Jake Gryz and Nima Shahbazi. Futility of a right to explanation, 2020. URL `http://ceur-ws.org/vol-2578/pie1.pdf`.

[78] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. arxiv preprint, 2019. URL `https://arxiv.org/abs/1909.03166`.

[79] Philipp Hacker and Jan-Hendrik Passoth. Varieties of AI explanations under the law. from the GDPR to the AIA, and beyond. In *xxAI – Beyond Explainable AI*, volume 13200 of *Springer eBook Collection*, pages 343–373. Springer International Publishing and Imprint Springer, 2022.

[80] Neal R. Haddaway, Matthew J. Grainger, and Charles T. Gray. Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research synthesis methods*, 13(4):533–545, 2022.

[81] Patrick Hall and Navdeep Gill. Debugging the black-box COMPAS risk assessment instrument to diagnose and remediate bias, 2017. URL `https://openreview.net/forum?id=r1iwhvj7z`.

[82] Patrick Hall, Navdeep Gill, and Nicholas Schmidt. Proposed guidelines for the responsible use of explainable machine learning. In *Robust AI in Financial Services Workshop at the 33rd Conference on Neural Information Processing Systems*, 2019.

[83] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul de Hert. Bridging the gap between ai and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1):72–85, 2022.

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

[85] Bernease Herman. The promise and peril of human evaluation for model interpretability. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[86] James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou. Fairness by explicability and adversarial SHAP learning. In *Machine Learning and Knowledge Discovery in Databases*, volume 12459 of *Springer eBook Collection*, pages 174–190. Springer International Publishing and Imprint Springer, 2021.

[87] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ACM Digital Library, pages 123–129. Association for Computing Machinery, 2019.

[88] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 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*, ACM Digital Library, pages 1–13. Association for Computing Machinery, 2019.

[89] Alexey Ignatiev, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva. Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming*, pages 846–867. Springer, Cham, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-58475-7_49.

[90] Aditya Jain, Manish Ravula, and Joydeep Ghosh. Biased models have biased explanations. arxiv preprint, 2020. URL https://arxiv.org/abs/2012.10986.

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

[92] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, pages 1–26, 2022.

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

[94] Thomas Kehrenberg, Myles Bartlett, Oliver Thomas, and Novi Quadrianto. Null-sampling for interpretable and fair representations. In *European Conference on Computer Vision*. Springer, Cham, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-58574-7_34.

[95] Barbara Kitchenham and Stuart M. Charters. Guidelines for performing systematic literature reviews in software engineering, 2007. URL https://www.researchgate.net/profile/barbara-kitchenham/publication/302924724_guidelines_for_performing_systematic_literature_reviews_in_software_engineering.

[96] René F. Kizilcec. 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*, ACM Digital Library, pages 2390–2395. ACM, 2016.

[97] Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC '19)*, 2019.

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

[99] Udo Kuckartz and Stefan Rädiker. *Analyzing qualitative data with MAXQDA: Text, audio, and video*. Springer eBooks Social Sciences. Springer, Cham, 2019. ISBN 9783030156718. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5926278.

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

[101] Himabindu Lakkaraju and Osbert Bastani. "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*, 2020.

[102] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:1–24, 2021.

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

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

[105] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–26, 2019.

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

[107] David Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, 2019. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403301`.

[108] Gerald S. Leventhal. What should be done with equity theory? In *Social Exchange*, pages 27–55. Springer, Boston, MA, 1980.

[109] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):1–45, 2020.

[110] Zachary C. Lipton. The mythos of model interpretability. *ACM Queue*, 16(3):31–57, 2018.

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

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

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

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

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

[116] Chuizheng Meng, Loc Trinh, Nan Xu, and Yan Liu. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports*, 12(7166):1–28, 2022.

[117] Loizos Michael. Machine coaching. In *Proceedings of the IJCAI 2019*, pages 80–86. International Joint Conferences on Artificial Intelligence, 2019.

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

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

[120] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016.

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

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

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

[124] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems*, 12(3):1–30, 2022.

[125] Nassih, Rym and Abdelaziz Berrado. 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*, 2020.

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

[127] Julian Nyarko, Sharad Goel, and Roseanna Sommers. Breaking taboos in fair machine learning: An experimental study. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, ACM Digital Library, pages 1–11. Association for Computing Machinery, 2021.

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

[129] Deepak P, Sanil V, and Joemon M. Jose. On fairness and interpretability. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[130] 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, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, 372, 2021.

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

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

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

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

[135] Alvaro Parafita and Jordi Vitrià. 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*, 2021.

[136] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 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*, ACM Digital Library, pages 1–15. Association for Computing Machinery, 2021.

[137] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *CHI Conference on Human Factors in Computing Systems*, ACM Digital Library, pages 1–22. Association for Computing Machinery, 2022.

[138] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. FAIR: Fair adversarial instance re-weighting. *Neurocomputing*, 476, 2022.

[139] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. Interpretable data-based explanations for fairness debugging. In *Proceedings of the 2022 International Conference on Management of Data*, ACM Digital Library, pages 247–261. Association for Computing Machinery, 2022.

[140] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

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

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

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

[144] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, Cham, 2018.

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

[146] Avi Rosenfeld and Ariella Richardson. Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019.

[147] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[148] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

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

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

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

[152] Jakob Schoeffer, Niklas Kuehl, and Isabel Valera. A ranking approach to fair classification. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. ACM, 2021.

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

[154] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. On the relationship between explanations, fairness perceptions, and decisions. arxiv preprint, 2022. URL `https://arxiv.org/abs/2204.13156`.

[155] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. On explanations, fairness, and appropriate reliance in human-AI decision-making, 2022. URL `https://arxiv.org/abs/2209.11812`.

[156] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 2022.

[157] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 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)*, 2022.

[158] Andrew D. Selbst and Solon Barocas. The intuitive appeal of explainable machines. *Fordham Law Review*, 87(3):1085–1139, 2018.

[159] William Seymour. Detecting bias: Does an algorithm have to be transparent in order to be fair? In *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems*, pages 2–8, 2018.

[160] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 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*, pages 166–172, 2020.

[161] Donghee Shin. User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4):541–565, 2020.

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

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

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

[165] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 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*, pages 1–28, 2022.

[166] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*, 24(1):1–13, 2022.

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

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

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

[170] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In *Advances in Neural Information Processing Systems*, volume 34, pages 62–75. Curran Associates, Inc, 2021.

[171] Kacper Sokol and Peter Flach. Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In *SafeAI@AAAI*, 2019.

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

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

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

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

[176] Cor Steging, Silja Renooij, and Bart Verheij. 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*, ACM Digital Library, pages 235–239. Association for Computing Machinery, 2021.

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

[178] Simone Stumpf, Lorenzo Strappelli, Subeida Ahmed, Yuri Nakao, Aisha Naseer, Giulia Del Gamba, and Daniele Regoli. Design methods for artificial intelligence fairness and transparency. In *Joint Proceedings of the ACM IUI 2021 Workshops*, 2021.

[179] Gero Szepannek and Karsten Lübke. Facing the challenges of developing fair risk scoring models. *Frontiers in artificial intelligence*, 4:1–9, 2021.

[180] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 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*, 2019.

[181] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. FairTest: Discovering unwarranted associations in data-driven applications. In *2nd IEEE European Symposium on Security and Privacy*, pages 401–416. IEEE, 2017.

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

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

[184] Lingraj S. Vannur, Balaji Ganesan, Lokesh Nagalapatti, Hima Patel, and M. N. Tippeswamy. Data augmentation for fairness in personal knowledge base population. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 143–152. Springer, Cham, 2021. URL `https://link.springer.com/chapter/10.1007/978-3-030-75015-2_15`.

[185] Carla Piazzon Vieira and Luciano Antonio Digiampietri. Machine learning post-hoc interpretability: A systematic mapping study. In *XVIII Brazilian Symposium on Information Systems*, ACM Digital Library, pages 1–8. Association for Computing Machinery, 2022.

[186] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*, pages 1–47, 2017.

[187] B. Wagner and A. S. d'Avila Garcez. Neural-symbolic integration for fairness in AI. In *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering*, 2021. URL `https://openaccess.city.ac.uk/id/eprint/26151/`.

[188] Robyn Repko Waller and Russell L. Waller. Assembled bias: Beyond transparent algorithmic bias. *Minds and Machines*, pages 1–30, 2022.

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

[190] Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, pages 1–63, 2022.

[191] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 7285–7292, 2020. URL `https://ojs.aaai.org/index.php/aaai/article/view/6220`.

[192] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-makingz. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM Digital Library, pages 1–14. Association for Computing Machinery, 2020.

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

[194] David S. Watson and Luciano Floridi. The explanation game: A formal framework for interpretable machine learning. In *Ethics, Governance, and Policies in Artificial Intelligence*, volume 144 of *Springer eBook Collection*. Springer International Publishing and Imprint Springer, 2021.

[195] Claes Wohlin. 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 Digital Library, pages 1–10. ACM, 2014.

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

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

[198] Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, and Finale Doshi-Velez. Incorporating interpretable output constraints in bayesian neural networks. In *34th Conference on Neural Information Processing Systems*, 2020.

[199] Hantian Zhang, Nima Shahbazi, Xu Chu, and Abolfazl Asudeh. 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*, 2021.

[200] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making — The causal explanation formula. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[201] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 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*. ACM, 2020.

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

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

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