Minds and Machines (2024) 34:9
https://doi.org/10.1007/s11023-024-09672-2

A Genealogical Approach to Algorithmic Bias

Marta Ziosi[1] · **David Watson**[2] · **Luciano Floridi**[3,4]

Received: 24 October 2023 / Accepted: 6 March 2024

Abstract
The Fairness, Accountability, and Transparency (FAccT) literature tends to focus on bias as a problem that requires *ex post* solutions (e.g. fairness metrics), rather than addressing the underlying social and technical conditions that (re)produce it. In this article, we propose a complementary strategy that uses genealogy as a constructive, epistemic critique to explain algorithmic bias in terms of the conditions that enable it. We focus on XAI feature attributions (Shapley values) and counterfactual approaches as potential tools to gauge these conditions and offer two main contributions. One is constructive: we develop a theoretical framework to classify these approaches according to their relevance for bias as evidence of social disparities. We draw on Pearl's ladder of causation (Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, 2000, Causality, 2nd edn. Cambridge University Press, Cambridge, 2009. https://doi.org/10.1017/CBO9780511803161) to order these XAI approaches concerning their ability to answer fairness-relevant questions and identify fairness-relevant solutions. The other contribution is critical: we evaluate these approaches in terms of their assumptions about the role of protected characteristics in discriminatory outcomes. We achieve this by building on Kohler-Hausmann's (Northwest Univ Law Rev 113(5):1163–1227, 2019) constructivist theory of discrimination. We derive three recommendations for XAI practitioners to develop and AI policymakers to regulate tools that address algorithmic bias in its conditions and hence mitigate its future occurrence.

Keywords Artificial Intelligence · Bias · Epistemology · Ethics · Explainability · Genealogy · Machine learning

✉ Marta Ziosi
marta.ziosi@sant.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[2] Department of Informatics, King's College London, Bush House, 30 Aldwych, London WC2B 4BG, UK

[3] Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, USA

[4] Department of Legal Studies, University of Bologna, Via Zamboni, 27, 40126 Bologna, Italy

Published online: 02 May 2024

1 Introduction

Risk prediction tools can increase decision efficiency in contexts such as credit, health, and criminal justice. They may bring more neutrality, countering subjective and prejudice-driven human judgment, and improve accuracy, resulting in more efficient and resource-effective decision policies (Barabas et al., 2018). However, for years, algorithmic tools have been criticised for reflecting and potentially exacerbating pre-existing biases (Barocas & Selbst, 2016; Citron & Pasquale, 2014). "Algorithmic bias", in this context, is generally taken to refer to cases in which "the model's predictive performance (however defined) unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class" (Mitchell et al., 2021, p. 1). This bias is also referred to as a model's "skewed performance" along one of these demographic axes.

In the Fairness, Accountability, and Transparency (FAccT) literature, bias tends to be characterized as a *problem* in its consequences, that is, an issue requiring an *ex post* solution. An example of this type of solution is fairness metrics, a set of measures that enable one to detect and adjust bias in a model. Algorithmic bias is rarely considered as *evidence* of the underlying social and technical conditions that (re)produce it—that is, as an issue requiring an *ex ante* solution. This tendency promotes the design of solutions *ex post*, by addressing the consequences, rather than (at least also) *ex ante*, by addressing the conditions. In this article, we seek to rebalance the overall strategy. We analyse explainable artificial intelligence (XAI) approaches with respect to their ability to gather evidence—note, not proof—of social disparities. We focus specifically on feature attribution approaches that rely on Shapley values and counterfactual approaches. These enable us to examine the relationship between protected characteristics such as race or gender and skewed performance.

Although the relation between explainability and fairness is key to approaching algorithmic bias as "evidence", it remains analytically vague. Additionally, some of the bias-relevant applications of feature attribution approaches tend to represent the role of protected characteristics in discriminatory outcomes unrealistically—e.g., as independent, intrinsic, and causal attributes. A complementary strategy is to approach bias *genealogically*. In this article, we use genealogy as a constructive, epistemic critique,[1] with a double role. Constructively, it allows us to explain algorithmic bias in terms of the conditions that give rise to it, *ex ante*. Critically, it helps explain algorithmic bias not in terms of a single origin ("cause"), but with respect to a broader set of social and technical conditions at play that (re)produce these disparities.

In this respect, we make two main contributions. We offer a theoretical framework to classify XAI approaches according to their relevance to gather evidence of social disparities. We take inspiration from Pearl's ladder of causation (2000, 2009)

[1] In philosophy, following the tradition of Nietzsche and Foucault, a genealogy is a form of historical critique, designed to overturn our norms by revealing their origins (Hill, 2016). We use the term more in its philological sense, to mean a constructive critique that looks at the conditions of possibility for a problem to address it successfully.

to characterize XAI approaches into observational, interventional and counterfactual approaches—namely, concerning their ability to detect (a) whether—and, if so, (b) how—a protected characteristic contributed to skewed performance, and (c) what can be done to change it. The goal is to consider these XAI methods not only as technical tools but as means to investigate and collect evidence about unfair differences in performance alongside protected characteristics. The second is to critique these methods concerning their ability to represent the role of protected characteristics in discriminatory outcomes. Drawing from Kohler-Hausmann's (2019) constructivist theory of discrimination, we question observational, interventional, and counterfactual XAI approaches concerning the *independence*, *responsibility,* and *epistemic* assumptions they make towards protected characteristics, respectively. The aim is to question XAI approaches in their ability to help capture salient aspects of discrimination.

The remainder of this article is structured as follows. In Sect. 2, we review the relationship between explainability and fairness. In Sect. 3, we present the genealogical approach to bias. In Sect. 4, we characterize XAI approaches concerning their relevance for fairness. In Sect. 5, we question their capacity to address algorithmic discrimination. Finally, we derive three main recommendations for XAI practitioners to develop and policymakers to regulate tools that address algorithmic bias in its conditions and thus mitigate its future occurrence.

2 Explainability and Fairness

Explainability can enhance fairness-relevant properties to different extents and on different levels. Explainability can enhance transparency (Abdollahi & Nasraoui, 2018), granted by the ability to see how a model has arrived at a discriminatory outcome. Additionally, explainability can increase or enable trust in a model (Dodge et al., 2019). It can help, for example, to determine if qualities relevant to algorithmic fairness (such as fairness metrics) are met (Doshi-Velez & Kim, 2017). Explainability can also enhance accountability, as it can provide explanations for AI-informed (un)fair decisions (Leben, 2023; Zhou et al., 2022). These properties concern fairness within the context of "responsible AI", AI that takes into account moral, and ethical considerations as well as social values (Adadi & Berrada, 2018).

At the same time, the relationship between explainability and fairness is not always positive. Explainability, for instance, can influence the *perception* of fairness. On that note, some highlight the risks for fairness that more explainability, and more reliance on it, can bring. Examples are the risks of "fairwashing" and of the rationalization, and potential justification, of some types of discrimination (Aivodji et al., 2019). Ananny and Crawford (2018) have highlighted how the ability to "see" a model does not equate to the ability to "govern" it nor "understand" it and, thus, to mitigate bias. Additionally, authors such as Barocas (2022) have recently highlighted the tensions between calls for simpler models to ensure transparency (and thereby facilitate algorithmic fairness), and the inconvenient fact that such models may be less able to satisfy some fairness demands (e.g. allowing for specific

parameter tweaks). This shows how the relation between fairness and explainability need not be positive and may require trade-offs.

Notwithstanding the ability of XAI methods to enhance or mitigate the potentially negative consequences of algorithmic bias, our focus here is different. Specifically, we are here interested in the potential of XAI's methods to gather evidence of the conditions that enable it. We see algorithmic bias as the *object*, and XAI approaches as a *means* to uncover and learn about the underlying conditions of social inequality.

Within the realm of explainability approaches, we focus specifically on feature attribution (e.g. SHAP Lundberg & Lee, 2017) and counterfactual approaches (e.g. Galhotra et al., 2021; Karimi et al., 2020, 2021). Among feature attribution approaches, this article focuses specifically on Shapley values,[2] which can estimate how input features contribute to performance biases (Begley et al., 2020). This is arguably the most popular feature attribution method, as it unifies several related methods and comes with axiomatic guarantees (Lundberg & Lee, 2017). However, various methods exist for computing Shapley values that may provide different attributions for the same prediction (Sundararajan & Najmi, 2020). Following Heskes et al. (2020), we provide a division into marginal, conditional, and interventional or causal approaches. We also examine counterfactual approaches. These explain what could have happened to an outcome had an input feature to a model been changed in a particular way (Barocas et al., 2020; Verma et al., 2020). Together with the latter, as we will present, these approaches can be easily interpreted through a causal ladder framework.

3 A Genealogical Approach to Bias

The above suggests that XAI methods can be relevant to explaining an algorithm's performance concerning estimating the contribution of protected characteristics. When this performance reveals disparities about gender or race, the approaches we consider can support the goal of explaining this performance in terms of the input variables that conditioned it. However, we need an overarching strategy for how to apply the XAI methods we consider towards that goal. In this article, we propose to adopt a genealogical approach to algorithmic bias.

Genealogy refers to a form of historical critique, designed to overturn social norms by revealing their origins (Hill, 2016). Here, we use the term in its philological sense, to mean a constructive critique that looks at the conditions of possibility of a problem to address it successfully. In our case, this refers to a constructive

[2] Shapley values were adapted from Shapley's (1951) foundational work in cooperative game theory, where the original goal was to quantify the contribution of individuals to a given coalition. In the XAI context, Shapley values represent the marginal contribution of a feature to the model's output when averaged over a specific distribution of all possible feature coalitions. They are the unique solution to the attribution problem satisfying certain desirable properties—e.g., linearity, symmetry, and efficiency. Concerning fairness, they can help identify variables that are drivers of unfair outcomes. They do so by allocating responsibility for observed disparities, defined through a specific measure of fairness, among the considered input variables of the model (Lundberg, 2018).

critique designed to understand algorithmic bias by focusing on the plural, dynamic, and contingent conditions for its possibility, and the potential of XAI methods to surface evidence of them. Specifically, a genealogical approach to algorithmic bias invites us to explain bias in terms of the conditions for its occurrence and understand its explanation not in terms of a single cause, but of gathering evidence on the set of conditions that produce it.

As cited above, past redlining divisions, differential access to healthcare, and disparate arrest practices towards people of colour represent some examples of what we mean by these "conditions". Historical segregation in US neighbourhoods, for instance, profoundly affected the residents' access to credit, health insurance, and education (Agyeman, 2021; Perrino, 2020). In turn, this created the conditions of poverty, unemployment, and past default history by which residents in these communities are considered "not worthy" of credit when zip codes are used to calculate the risk of default. We aim to express this genealogical approach by adopting both a constructive and a critical stance, taking inspiration from Pearl's ladder of causation (2000, 2009) and Kohler-Hausmann's (2019) constructivist theory of discrimination, whose contribution will become more evident in the following sections.

4 XAI Approaches as Questions

By taking inspiration from Pearl's ladder of causation (2000, 2009), we provide an ordering principle for XAI approaches to clearly distinguish between their utility for fairness along three levels—specifically, their differential ability to "see", "govern", and "understand" what influences skewed performance. Generally, this should help clarify the vague relationship between explainability and fairness. In the specific case of the feature attribution and counterfactual approaches that this article focuses on, this can help answer the following questions: (1) Is a protected characteristic unfairly associated with outcomes? (2) Would intervening to alter a protected characteristic directly affect outcomes? (3) Given observed values for protected characteristics and outcomes, would a hypothetical intervention to alter a subject's protected characteristic have changed the outcome? We propose so-called "observational" approaches as relevant for procedural fairness, "interventional" approaches for consequential recommendations, and "counterfactual" approaches for algorithmic recourse.

4.1 What Bias: Observational Approaches for Procedural Fairness

We refer to marginal and conditional feature attribution methods as "observational approaches", as they can help observe whether a protected characteristic is unfairly associated with an outcome. Marginal variable importance measures estimate the importance of features, assuming that these are independent of each other. An example is given by Datta et al. (2016)'s Quantitative Input Influence (QII) method, where Shapley values are used to calculate the average marginal influence of input features. Another example is provided by Štrumbelj

and Kononenko's (2014) work. They use marginal Shapley values to develop a sensitivity analysis-based method to estimate individual feature contributions. Even though common, this assumption of independence might lead to incorrect or counterintuitive explanations when the features are, in fact, highly correlated. Additionally, it allows these methods to represent only the direct effects of variables.

This motivates the use of conditional variable importance measures. These can represent indirect effects, and estimate importance by conditioning on a variable. Aas et al. (2021) propose conditioning strategies to compute more accurate Shapley values. Another example is Frye et al.'s (2020) so-called asymmetric Shapley values. They are called "asymmetric" because, when computing Shapley values, they restrict the possible permutations of the features to those consistent with a partial causal ordering. They then apply conditioning by observation to check that their explanations respect the multivariate distribution of the data. Thus, they check that they do not produce misleading or nonsensical explanations because of feature dependence.

Overall, these observational approaches make it possible to check whether a protected characteristic, such as race or gender, is unfairly associated with an outcome. This is relevant for ensuring procedural fairness, i.e. the fairness of the decision-making process (Grgić-Hlača et al., 2018). In the example of credit risk assessment, it would allow one to see whether, for example, gender or race was considered in arriving at a negative assessment for a loan. In this respect, "unfairness" would amount to direct discrimination and be illegal (Prince & Schwarcz, 2019). Procedural fairness is related to commitments to values such as accountability and transparency (Rueda et al., 2022). Additionally, it allows the fulfillment of requests such as compliance in finance or due process in law.

These considerations suggest that these measures could be relevant for fairness in an algorithmic context for tasks such as auditing [reference anonymised]; specifically, ethics-based auditing (EBA; Mökander et al., 2021; reference anonymised). EBA refers to "a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms" (Mökander et al., 2021, p. 2). According to Mökander et al. (2021), EBA can "contribute to good governance by promoting procedural regularity and transparency" (p. 16). The feature importance approaches mentioned here can contribute to assessing whether a protected characteristic played a role in the decision-making process.

In this respect, while both marginal and conditional approaches are concerned with answering the question above, they might be differentially relevant for procedural fairness. Authors like [reference anonymised] suggest that the former can provide insights into model mechanics, while the latter is more informative about the underlying data-generating process. Accordingly, marginal measures could help shed light on discrimination at the level of the model. This could be relevant to EBA when reviewing source code audit (Mökander et al., 2021). Conditional measures could help shed light on discrimination at the system level. In this respect, a model could play an instrumental role, and conditional measures could be useful to check for procedural fairness in more complex decisional settings such as institutions for credit, justice, etc.

4.2 How Bias: Interventional Approaches for Consequential Recommendations

Interventional approaches can help investigate whether intervening to alter a protected characteristic directly affects outcomes. Beyond showing whether a feature is correlated with a discriminatory outcome, these approaches allow us to tell how (e.g. the extent to which) an intentional change in one feature changes the outcome. So-called interventional or causal Shapley values can help answer this question, as they are designed to capture causal contributions (Heskes et al., 2020). They do so by quantifying the effects of each input on a model's output in accordance with a user-supplied causal graph. Examples include do-Shapley values (Jung et al., 2022) and Shapley flow (Wang et al., 2021).

Unlike marginal observational approaches, most of these approaches consider the relations between input features (Heskes et al., 2020). They do not assume independence between them. They tend to do so by relying on a causal representation of the model or the "world" through a causal diagram. This entails making explicit some assumptions about the features considered in a model. Such a representation can help understand the interaction of the input features within a model or a system, and to reason about potential interventions (Wang et al., 2021). When modelled as a causal driver of both re-arrests and anti-law enforcement resentment ("antisocial cognition"), for example, intensive policing could be reasoned about as a site to intervene on (e.g. to be reduced) to mitigate these outcomes.

Regarding fairness, this can potentially respect and enhance one's agency. The ability for someone to exercise their agency is close to values such as self-determination and autonomy (Christman, 2020). Interventional approaches could do so in the form of consequential recommendations. Given a negative outcome, consequential recommendations provide the minimum intervention required to obtain a better result (Karimi et al., 2020). For example, they might suggest how much a credit applicant needs to increase their credit score or income to raise their chances of receiving a loan. However, since most protected characteristics cannot be changed, these recommendations suggest interventions on so-called "intervenable" factors, such as income in credit risk assessment, employment in crime risk assessment or nutrition in health risk assessment.

4.3 Why Bias: Counterfactual Approaches for Algorithmic Recourse

Given observed values for protected characteristics and outcomes, these approaches can suggest whether a hypothetical intervention to alter a subject's protected characteristic would have changed the outcome. Examples are provided by Galhotra et al. (2021) and Karimi et al., (2020, 2021). Most notably, Galhotra et al. (2021) propose "probabilistic contrastive counterfactuals" which not only help quantify the direct and indirect effects of a feature on outcomes, but also provide actionable recourse to individuals negatively affected by such an outcome.

These methods can allow users to check whether a protected characteristic (e.g. race or gender) was the cause of a specific discriminatory outcome, and what can be done to change that outcome. Beyond interventional approaches, counterfactual

ones allow one to know not only how to act, but also to understand what brought about a discriminatory outcome. Additionally, interventional approaches base their recommendations on the *consequences* they can bring about. They are forward-looking. By contrast, counterfactual approaches can base recommendations on what *caused* a discriminatory outcome. They are backward-looking. One is concerned with improving outcomes, the other with reversing unfavourable outcomes. In that respect, their relevance to fairness is still related to agency, but more closely to "recourse".

In law, recourse refers to actions that individuals or corporations undertake to remedy unfair or unfavourable legal outcomes (Wallin, 1992). Algorithmic recourse refers to "the systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios" (p. 284) (Venkatasubramanian & Alfano, 2020). For example, a rejected loan applicant who is a woman can argue for recourse if there exists a positive counterfactual instance to hers; an applicant who is similar or "close" to her in every other feature but for gender, and who was granted a loan. Karimi et al. (2021) consider that algorithmic recourse is met when a candidate is provided both with an explanation of why the loan was rejected and offered recommendation(s) on how to obtain the loan in the future (Karimi et al., 2021). Algorithmic recourse, they claim, is achieved when one can "can *understand* and accordingly *act* to alleviate an unfavourable situation" (p. 2) (Karimi et al., 2021). Given that these explanations are formed by looking at an opposite outcome in a unit that is the same or similar *but for* a protected feature, these explanations are often referred to as contrastive explanations (Galhotra et al., 2021; Karimi et al., 2021). These formulate explanations in terms of explaining why this outcome rather than another ("the opposite") happened.

5 Questioning XAI Approaches

What does it mean, however, for a protected feature such as gender or race to "be associated with", "alter" or "cause" a discriminatory outcome? This section takes inspiration from Kohler-Hausmann's (2019) constructivist theory of discrimination to question how these XAI methods approach it with respect to protected features.

5.1 From Procedural Explanations to Evidential Observations

From a statistical standpoint, the main objection to using observational approaches for procedural fairness arises from their assumption of feature independence. Marginal Shapley values ignore the fact that a change in one input feature may cause a change in another. If protected characteristics present spurious correlations with the discriminatory outcome via another, possibly unobserved, feature this will produce misleading explanations (Heskes et al., 2020; Nabi & Shpitser, 2018). Conditional Shapley values recognize the presence of other features, and how this can influence the contribution of other features under consideration. However, they do not usually rely on a causal representation (e.g. a causal graph) of the relationship between

characteristics. Similarly to some interventional approaches, they might not distinguish intermediate outcomes from covariates (Greiner, 2008) and produce unreliable explanations.

From a critical standpoint, we should ask what it means for a protected characteristic such as race or gender to be independent of other input features. We claim that the problem with assuming independence is that it assumes that protected features can be represented as discrete units, existing in isolation rather than in relation to a host of other variables or features.

The independence assumption comes with important conceptual repercussions. In reality, one cannot separate being a woman or being a person of colour from one's socioeconomic circumstances, at least in societies where gender or racial inequalities are present (Hu, 2019). Even when this dependence is acknowledged by conditional approaches, i.e., by "controlling for" or "conditioning on" gender, it should be understood in relation to the entire system of other features within which it is embedded. To attribute a role to gender in a credit decision without considering how variables such as income, marital status, and education relate to each other is to misunderstand or even deny the very role that gender plays in credit settings. One is not classified as less likely to pay back a loan just because they are a woman, but because of how that "comes with", and thus influences, a host of other input features, i.e., lower income because of the gender pay gap.

These approaches are thus more appropriate to *observe* whether gender played a role, rather than to explain *which* role gender played vis-à-vis other features. While this aligns with the aims of procedural fairness, these considerations should serve as a caveat to avoid using these approaches beyond their means. It is essential to avoid using the contributions they estimate as full explanations, but rather as mere evidence of a potential link with gender, which needs to be further tested with, for example, causal inference methods or by engaging with the individuals or communities affected. We thus suggest referring to explanations produced through these approaches as "evidential observations", explanations that may trigger further investigation into the conditions that brought about a discriminatory outcome.

5.2 From Consequential to Constructive Recommendations

One objection that can be brought to consequential recommendations obtained through interventional approaches is precisely that they do not rely on causes, but on whatever brings about the best consequences. As a result, the actions they offer can be either inconsistent or have little to do with discrimination. For example, recommendations tailored towards making rejected applicants more likely to be accepted for a loan might require them to change their race. As this is impossible (and wrong), other input features on which an individual can more plausibly intervene are often used instead. In this case, recommendations can provide a relatively sensible suggestion, such as changing one's job. However, if one's rejection was, in fact, a result of (racial) discrimination, this does little to resolve the original injustice.

We ought to ask, what is problematic about "intervening on" a protected characteristic such as race or gender? This request, we argue, entails an assumption

about responsibility. Namely, it ascribes to the person with the protected feature the responsibility to change their situation. However, protected features are traits *of* the unit. Not only can one not practically intervene on or change their race, but also, normatively, one should not be held responsible for it. This is specifically relevant considering our previous claim; that interventional approaches can be useful for fairness with regard to enhancing one's agency. When developing consequential recommendations, it is crucial that the responsibility for changing an outcome rests on the people who designed and deployed the algorithm (e.g., developers or providers), not the end users subjected to their results.

As such, any consequential recommendation that focuses on the effect of protected characteristics on a discriminatory outcome should not simply rely on another feature that can be "intervened on". It should instead provide a reason for the developers to reflect on how its model represents these protected characteristics. Given that interventional and causal Shapley values usually rely on a causal representation of the model or the "world" through a causal graph, these approaches can promote fairness when developers use them to understand how protected features such as race or gender are represented in their model alongside a set of other features (Hu, 2019). This is especially relevant given that causal graphs encode direct and indirect associations. Thus, they allow one to reason about positive and negative contributions to skewed performance.

As Kohler-Hausmann (2019) suggests, protected features do not have causal effects so much as structural properties: they are *embedded* within structures, whether social systems or algorithmic models, influencing and constructing their meaning and role. As such, these approaches could be used to provide constructive recommendations to developers on how to change the way gender or race are represented in the model vis-à-vis other input features. For example, this could be done by realizing that some variables seemingly unrelated to race, e.g. zip codes, are proxies for it in their model and how they could reconfigure this relation to change it. Rather than bringing about desired consequences for and from users, these approaches can be better framed as constructive recommendations from developers toward users.

5.3 From Contrastive to Constitutive Explanations

Regarding counterfactual approaches, one could contend that, even though theoretically sound, they are often too demanding to realize in practice (Verma et al., 2020). Computing counterfactuals generally requires not just a causal graph but also knowledge of the structural equations that govern the relationships between nodes. As we noted previously, some XAI approaches that rely on this intuition overcome these limits by framing the search for the counterfactual as an optimization problem. The counterfactual is found by characterizing a notion of distance that allows us to identify the nearest hypothetical point, which is classified differently from the one considered (Wachter et al., 2017).

While this practical solution reduces the epistemic demands of counterfactuals, we should ask what it means for gender or race to 'cause' a discriminatory outcome. The solution provided above relies on the assumption that the two "similar"

or "nearest" units are the same *but for* the protected features. For a protected feature to be causal, as we suggested before, would mean that it is the only factor by which these two similar units differ and that, given that these lead to two different outcomes, the protected feature must be the cause. However, as Kohler-Hausmann (2019) suggests, a counterfactual unit is not the same *but for* the protected feature. It is a different unit precisely *because of* the protected feature. The epistemic assumption that contrastive explanations make is that gender or race are "causes" rather than that they constitute or characterize different units or "worlds".

In this respect, one should talk about constitutive rather than counterfactual explanations. To say that a given system has a causal effect because of how it is constituted is to suggest that if one changed parts of that system, it would have a different causal effect. However, by the logic above, it would also be a different system. In this sense, a constitutive explanation attempts to explain different outcomes by pointing not to the "cause" but to the parts that constitute these different units. With relevance to fairness, explaining different outcomes concerning these parts and their organization can help shed light on the conditions by which similar individuals are treated differently. For example, given two similar individuals with different credit outcomes, a constitutive explanation would entail naming the features by which these individuals are considered similar precisely as what makes their different treatment unfair. Going back to the aim of this article, it would help approach bias *ex ante* by focusing on the conditions that constitute it, rather than only *ex post* by focusing on fixing the consequences.

6 Recommendations for AI Policy and Governance

This article argues that the Fairness, Accountability, and Transparency (FAccT) literature tends to focus on bias as a problem that requires ex post solutions, e.g. fairness metrics, rather than also addressing the underlying social and technical conditions that (re)produce it. It proposes a complementary strategy that uses genealogy as a constructive, epistemic critique to explain algorithmic bias in terms of the conditions for its possibility. In this respect, the article has focused on XAI feature attribution approaches (Shapley values) and counterfactuals as potential tools to shed light on these conditions.

Given the considerations above, we conclude with three recommendations that can be useful for XAI practitioners (researchers, developers, and providers of XAI tools) when developing or deploying the XAI approaches and AI policymakers when regulating AI with fairness in mind.

(1) The relevance of explainability for fairness should be explicitly articulated and integrated in AI development and regulation.

Section 2 reports an unclear, and sometimes counterproductive relationship between XAI approaches and bias. This is backed by an increasing number of

libraries, such as IBM360[3] and WhatIf tools,[4] that offer explainability and fairness tools without providing specific guidance about which tools can best tackle which aspect of bias. IBM360, for example, provides two entirely separate libraries for fairness and explainability: AI Fairness360 and AI Explainability360. This leaves open the possibility that these XAI approaches may be mistakenly or deceitfully employed in their application to fairness problems; for instance, that an XAI method relevant for procedural fairness is used beyond its capacity to detect whether a protected feature played a role in the outcome, to investigate *which* role it played. As we claimed in this article, when developing and deploying XAI approaches, XAI practitioners should think about them concerning the fairness-relevant questions they can answer and the fairness-relevant solutions that they help identify; for example, as presented in this paper, whether they are useful for procedural fairness or to enhance or protect one's agency. Additionally, they should provide clear instructions about how their approaches can be used in concert with others and list their limitations and strengths concerning the fairness-relevant purpose they can play. It is also crucial that AI regulatory initiatives introduce measures that recognize and promote the coherence and complementarity of the properties of XAI methods for fairness. Without such recognition, we might miss out on the opportunities they offer but also enhance the risks entailed by their misuse. An example can be found in how one of the latest amendments to the AI Act weakened the fairness-relevant potential of explainability by shifting the focus on ensuring oversight and traceability rather than empowering end-users rights to explanations (Nannini et al., 2023).

(2) The responsibility to act on discriminatory outcomes should not lie exclusively with discriminated users.

XAI approaches should not only be designed for discriminated users seeking advice or recourse after being subjected to discrimination. Research suggests that many users might not even know that they are interacting with an algorithm, let alone that they have been discriminated against [reference anonymised]. It is crucial that XAI practitioners develop methods that not only help recognize instances of discrimination, but also provide constructive explanations on how to address their negative impact. These explanations could suggest interventions that AI providers can undertake to prevent discrimination or to intervene on it once it occurs (Karimi et al., 2021). This would not only help protect consumers from discrimination but also help AI providers prevent future liability claims under upcoming legislative proposals (Hacker, 2022). AI regulatory initiatives increasingly rely on ensuring compliance through risk management, audits and certification (Roberts et al., 2023). Additionally, it has been suggested that external validation of models by trusted third parties can ensure the reproducibility of results and surface biases (Haibe-Kains et al., 2020). Interventional XAI approaches could be used to provide

[3] AI Explainability 360 (https://aix360.mybluemix.net/) and AI Fairness 360 (https://aif360.mybluemix.net/).

[4] https://pair-code.github.io/what-if-tool/.

feedback to AI providers in the form of constructive recommendations from third-party audits [reference anonymised] to help them adhere to AI regulations.

(3) Explanations of discriminatory outcomes should name conditions, rather than just the main cause(s) of discrimination.

Hu (2019) suggests that causal graphs can be useful to explicitly state the assumptions one makes about a problem and to examine the social biases at play. Similarly, causal graphs could help represent the features that constitute a discriminatory outcome, and how they relate to each other. This approach would provide explanations that identify the primary factor responsible for discrimination and enhance our comprehension of how it relates to the set of similar conditions that contribute to an unjustifiably different outcome and how these conditions relate to one another. As mentioned, this can help shed light on the conditions by which similar individuals are treated differently. While research proposes the potential use of some fairness metrics to create a prima facie case for discrimination (Wachter et al., 2021), XAI approaches could go further and help provide a richer understanding of systemic discrimination. This is especially relevant in light of new regulatory initiatives such as the California Racial Justice Act (2020), which allows the use of statistical evidence in letting people charged with (or convicted of) a crime raise issues of racial bias and discrimination. As the law relies on a counterfactual intuition[5] to prove the existence of racial disparities, counterfactual approaches could provide a richer picture of what constitutes these disparities. As suggested in Sect. 5.3, this could be done by naming the features by which these two ethnically different yet similar individuals receive a different outcome as what makes their differential treatment one firmly rooted in systemic racism.

7 Conclusion

By shifting the focus to the conditions for rather than the consequences of discriminatory outcomes, this article hopes to emphasize the importance of understanding and preventing algorithmic discrimination. The genealogical approach proposed here, both in its constructive and critical components, can help tailor the application of XAI approaches not only to "see" discrimination but also to "govern" and "understand" its workings (Ananny & Crawford, 2018). At the same time, it can help us by recognising these approaches' limitations in representing and addressing algorithmic discrimination. Pragmatically, we have provided a set of policy recommendations by which both these constructive and critical components can be integrated into AI development and regulation.

Significantly, we also recognize that XAI approaches can only go so far in matters of discrimination. They should be envisioned as part of a more comprehensive strategy. Thus, in this article, we evaluate them in their potential to support, rather than

[5] Racial disparity can be suggested when a longer or more severe sentence was imposed on the defendant than was imposed on other similarly situated individuals convicted of the same offense.

supplant, approaches to address algorithmic discrimination. Thinking with a genealogical approach in mind, future research could explore how these XAI approaches could be used in concert with qualitative and contextual efforts to (re)construct how historical disparities emerge and are reproduced in AI systems. For example, their ability to surface hints of the conditions that enable discrimination could be corroborated through qualitative data and the participation of the communities involved to build a more comprehensive and accurate picture.

Funding This study was funded by the Centre for Digital Ethics (CEDE) of Bologna University.

Declarations

Conflict of interest The authors have no financial or proprietary interests in any material discussed in this article.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence, 298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

ACLU California Action. (2020). *AB 256*. ACLU California Action. https://aclucalaction.org/bill/ab-256/

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In J. Zhou & F. Chen (Eds.), *Human and machine learning: Visible, explainable, trustworthy and transparent* (pp. 21–35). Springer. https://doi.org/10.1007/978-3-319-90403-0_2

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agyeman, J. (2021, March 9). *How urban planning and housing policy helped create 'food apartheid' in US cities*. The Conversation. http://theconversation.com/how-urban-planning-and-housing-policy-helped-create-food-apartheid-in-us-cities-154433

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. In *Proceedings of the 36th international conference on machine learning*, 2019 (pp. 161–170). https://proceedings.mlr.press/v97/aivodji19a.html

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. arXiv:1712.08238 [Cs, Stat]. http://arxiv.org/abs/1712.08238

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020 (pp. 80–89). https://doi.org/10.1145/3351095.3372830

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. arXiv: 2010.07389 [Cs, Stat]. http://arxiv.org/abs/2010.07389

Christman, J. (2020). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Citron, D. K., & Pasquale, F. A. (2014). *The scored society: Due process for automated predictions* (SSRN Scholarly Paper ID 2376209). Social Science Research Network. https://papers.ssrn.com/abstract=2376209

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 2016 (pp. 598–617). https://doi.org/10.1109/SP.2016.42

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, 2019 (pp. 275–285). https://doi.org/10.1145/3301275.3302310

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608 [Cs, Stat]. http://arxiv.org/abs/1702.08608

Frye, C., Rowat, C., & Feige, I. (2020). Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Advances in neural information processing systems*, 2020 (Vol. 33, pp. 1229–1239). https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html

Galhotra, S., Pradhan, R., & Salimi, B. (2021). *Explaining black-box algorithms using probabilistic contrastive counterfactuals* (arXiv:2103.11972). https://doi.org/10.48550/arXiv.2103.11972

Greiner, D. J. (2008). Casual inference in civil rights litigation. *Harvard Law Review, 122*, 533.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018 (Vol. 32(1), Article 1). https://doi.org/10.1609/aaai.v32i1.11296

Hacker, P. (2022). The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4279796

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., Broderick, T., Hoffman, M. M., Leek, J. T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T.,…,Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature, 586*(7829), 7829. https://doi.org/10.1038/s41586-020-2766-y

Heskes, T., Sijben, E., Bucur, I. G., & Claassen, T. (2020). *Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models* (arXiv:2011.01625). https://doi.org/10.48550/arXiv.2011.01625

Hill, R. K. (2016). Genealogy. In *Routledge encyclopedia of philosophy* (1st ed.). Routledge. https://doi.org/10.4324/9780415249126-DE024-1

Hu, L. (2019). Disparate causes, Pt. I. *Phenomenal World*. https://www.phenomenalworld.org/analysis/disparate-causes-i/

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Bloebaum, P., & Bareinboim, E. (2022). On measuring causal contributions via do-interventions. In *Proceedings of the 39th international conference on machine learning*, 2022 (pp. 10476–10501). https://proceedings.mlr.press/v162/jung22a.html

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2021). *A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects* (arXiv:2010.04050). arXiv. http://arxiv.org/abs/2010.04050

Karimi, A.-H., Schölkopf, B., & Valera, I. (2020). *Algorithmic recourse: From counterfactual explanations to interventions* (arXiv:2002.06278). https://doi.org/10.48550/arXiv.2002.06278

Kohler-Hausmann, I. (2019). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review, 113*(5), 1163–1227.

Leben, D. (2023). Explainable AI as evidence of fair decisions. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2023.1069426

Lundberg, S. (2018). *Explaining quantitative measures of fairness—SHAP latest documentation*. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Explaining%20quantitative%20measures%20of%20fairness.html

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 2017 (Vol. 30). https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*(1), annurev-statistics-042720-125902. https://doi.org/10.1146/annurev-statistics-042720-125902

Mokander, J., & Floridi, L. (2021). *Ethics-based auditing to develop trustworthy AI*. SSRN Scholarly Paper ID 3788841. Social Science Research Network. https://papers.ssrn.com/abstract=3788841

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics, 27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

Nabi, R., & Shpitser, I. (2018). *Fair inference on outcomes* (arXiv:1705.10378). http://arxiv.org/abs/1705.10378

Nannini, L., Balayn, A., & Smith, A. L. (2023). *Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK* (arXiv:2304.11218). https://doi.org/10.48550/arXiv.2304.11218

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Perrino, J. (2020, July 2). *"Redlining" and health indicators: Decisions made 80 years ago have health consequences today.* NCRC. https://ncrc.org/redlining-and-health-indicators-decisions-made-80-years-ago-have-health-consequences-today/

Prince, A. E. R., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review, 105*, 1257.

Roberts, H., Ziosi, M., Osborne, C., Saouma, L., Belias, A., Buchser, M., Casovan, A., Kerry, C., Meltzer, J., Mohit, S., Ouimette, M.-E., Renda, A., Stix, C., Teather, E., Woodhouse, R., & Zeng, Y. (2023). *A comparative framework for AI regulatory policy*. https://ceimia.org/wp-content/uploads/2023/02/Comparative-Framework-for-AI-Regulatory-Policy.pdf

Rueda, J., Delgado, J., Parra Jounou, I., Hortal Carmona, J., Ausín, T., & Rodríguez-Arias, D. (2022). "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI and Society*. https://doi.org/10.1007/s00146-022-01614-9

Shapley, L. S. (1951). *A value for N-person games*. RAND Corporation. https://www.rand.org/pubs/papers/P295.html

Solon, B. (Director). (2022, August 19). *SRA22 Day 3—Keynote talk with Solon Barocas*. https://www.youtube.com/watch?v=Ft5rK1tTYyw

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sundararajan, M., & Najmi, A. (2020). The many Shapley values for model explanation. In *Proceedings of the 37th international conference on machine learning*, 2020 (pp. 9269–9278). https://proceedings.mlr.press/v119/sundararajan20b.html

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020 (pp. 284–293). https://doi.org/10.1145/3351095.3372876

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv:2010.10596 [Cs, Stat]. http://arxiv.org/abs/2010.10596

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*(2), 841–888.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). *Bias preservation in machine learning: The legality of fairness metrics under EU Non-Discrimination Law* (SSRN Scholarly Paper ID 3792772). Social Science Research Network. https://doi.org/10.2139/ssrn.3792772

Wallin, D. E. (1992). Legal recourse and the demand for auditing. *The Accounting Review, 67*(1), 121–147.

Wang, J., Wiens, J., & Lundberg, S. (2021). *Shapley flow: A graph-based approach to interpreting model predictions* (arXiv:2010.14592). https://doi.org/10.48550/arXiv.2010.14592

Zhou, J., Chen, F., & Holzinger, A. (2022). Towards explainability for AI fairness. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller & W. Samek (Eds.), *xxAI—Beyond explainable AI: International workshop, held in conjunction with ICML 2020: Revised and extended papers*, July 18, 2020, Vienna, Austria (pp. 375–386). Springer. https://doi.org/10.1007/978-3-031-04083-2_18