Forthcoming in *Feminist Philosophy Quarterly*

Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy

Linus Ta-Lun Huang, Society of Fellows in the Humanities, University of Hong Kong, Hong Kong

ORCID: 0000-0003-2469-8944

Hsiang-Yun Chen, [1] Institute of European and American Studies, Academia Sinica, Taiwan

ORCID: 0000-0002-9106-7846

Ying-Tung Lin, Institute of Philosophy of Mind and Cognition, National Yang Ming Chiao Tung University, Taiwan

ORCID: 0000-0001-5974-7860

Tsung-Ren Huang, Department of Psychology and CRETA, National Taiwan University, Taiwan

ORCID: 0000-0003-4396-7943

Tzu-Wei Hung, Institute of European and American Studies, Academia Sinica, Taiwan

ORCID: 0000-0003-4411-8038

Abstract

Artificial intelligence (AI) systems are increasingly adopted to make decisions in domains such as business, education, health care, and criminal justice. However, such algorithmic decision systems can have prevalent biases against marginalized social groups and undermine social justice. Explainable artificial intelligence (XAI) is a recent development aiming to make an AI system's decision processes less opaque and to expose its problematic biases. This paper argues against *technical XAI*, according to which the detection and interpretation of algorithmic bias can be handled more or less independently by technical experts who specialize in XAI methods. Drawing on resources from feminist epistemology, we show why *technical XAI* is mistaken. Specifically, we demonstrate that the proper detection of algorithmic bias requires relevant interpretive resources, which can only be made available, in practice, by actively involving a diverse group of stakeholders. Finally, we suggest how feminist theories can help shape *integrated XAI*: an inclusive social-epistemic process that facilitates the amelioration of algorithmic bias.

Keywords: algorithmic bias, explainable AI, feminist epistemology, situated knowledge, epistemic injustice

[1] Corresponding author.

1. Introduction

Imagine that you are a human resource officer tasked with recruiting new talent. Your company makes use of a state-of-the-art AI recruitment system that adopts a deep neural network model to predict a job candidate's "hireability score" based on over a hundred features, such as age, gender, race, personality traits, education level, years of relevant experience, and more.[2] Very roughly, the algorithm picks out statistical regularities that have been relevant in the past and makes predictions as to candidates' suitability on the basis of prior hiring decisions.[3] One of your responsibilities is to ensure hiring is done in an unbiased manner, and you see a disconcerting pattern: those recommended by the AI system seem unrepresentative of those who apply. Despite having a diverse applicant pool, the AI system appears to favor people who are homogeneous in terms of race and gender. Of course, this alone is not conclusive evidence that the recruitment system is biased. Other reasons, such as differences in qualification, may explain the results. However, because the recruitment system's decision criteria are implicitly embedded in its complex structure, it is extremely difficult for you to see or understand the reasons for its decisions and to figure out whether the process is indeed biased.[4]

To understand how an algorithmic decision is made, and to determine whether bias exists in a complex and opaque AI decision system, a new research program—explainable artificial intelligence (XAI)—was developed (Nunes and Jannach 2017; Adadi and Berrada 2018; Lipton 2018; Gunning and Aha 2019; Arrieta et al. 2020;). To a first approximation, XAI seeks to help humans understand and predict how AI systems generate decisions. Specifically, XAI helps identify significant factors that drive algorithmic decisions and offers comprehensible explanations to relevant stakeholders (Arrieta et al. 2020). Turning back to our example, it can explain why a certain candidate was, or was not, recommended for hire. The explanation

[2] We bracket debates on whether race and gender are real categories and work under the assumption that people are racialized and gendered according to current (unjust) social practices.

[3] More specifically, the AI system is exposed to historical data articulating applicants' features and prior hiring decisions during model training, "learning" the statistical regularities between these features and hiring decisions by rewiring connectivity among artificial neurons. At the end of the model's training phase, the AI system forms a processing stream that maps features of a candidate to a score, and then to a recommendation.

[4] Scientists apply a learning algorithm to a dataset to build a predictive model that can work as an AI decision system (Fazelpour and Danks 2021). Because the discussion about bias in AI decision systems is mostly about the predictive model, we use the terms "algorithm," "model," and "AI system" interchangeably. Additionally, we will be concerned with algorithmic bias in the broad sense, including any bias that is ultimately encoded in the predictive model, regardless of how it is introduced during the machine learning processes (Hellström, Dignum, and Bensch 2020).

often comes in the form of listing the most important features driving the decision—for instance, recommending candidate X because of her teamwork skills and relevant experiences.

Our focus is on the utilization of XAI for bias detection and reduction (Arrieta et al. 2020; Langer et al. 2021; Mohseni, Zarei, and Ragan 2021). Because the use of algorithms is increasingly common in many domains, including hiring, criminal justice, health care, and education, there are growing concerns that bias in such systems will impede social justice by perpetuating or even exacerbating existing inequalities (e.g., O'Neil 2016; Benjamin 2019; Noble 2018; D'Ignazio and Klein 2020). As measures for debiasing are in high demand, a thoroughgoing analysis of XAI and its promises is necessary.

We ask what XAI *ought* to be in order for it to help us combat bias in AI systems.[5] In particular, we will argue against a naive understanding—which we call *technical XAI*—according to which the detection and interpretation of algorithmic bias can be handled more or less by technical experts who specialize in XAI. Drawing from resources in feminist philosophy, we will argue that technical XAI is misguided. A better way to conceptualize and implement XAI for bias reduction—a view that we term *integrated XAI*—should replace technical XAI.

Our paper runs as follows. Section 2 reviews the current state of the art of XAI, defining and illustrating how technical XAI is supposed to help reduce algorithmic bias. Section 3 introduces insights from feminist epistemology, including feminist philosophy of science, situated knowledge, and epistemic injustice. Section 4 details the argument against technical XAI and motivates the move toward integrated XAI. Finally, section 5 makes the case for, and suggests ways to build, integrated XAI, drawing on feminist theoretical resources.

2. How XAI Can Reduce Bias

In this section, we articulate technical XAI (hereafter, TXAI), a technological solutionist approach to XAI for bias reduction. We will then illustrate how TXAI is supposed to extract relevant information from AI systems for the purpose of bias detection by reference to a model we developed and ran ourselves.

[5] Our project is closely related—but, strictly speaking, not equivalent—to what Haslanger (2012) describes as an *ameliorative* inquiry. Haslanger aims to ameliorate concepts (e.g., race and gender) for a certain legitimate purpose. Currently there are different ideas and methods associated with XAI; we ask how we *ought* to conceptualize and implement XAI for the purpose of bias reduction.

An AI system is biased if its output, procedure, and so on, deviate from a relevant (moral, epistemic, or social) norm (Fazelpour and Danks 2021). In this paper, we focus on morally problematic cases of procedural biases. Procedural bias occurs when (i) a score returned by an AI system for a case (e.g., in a hiring system) is influenced by a "class variable" (a variable that defines a specific attribute for a class), and (ii) in the most representative and restrictive case of procedural bias, this class variable is used by the model to generate the score directly. Procedural bias in this restrictive sense can be morally problematic because it can constitute disparate treatment (Glymour and Herington 2019). Disparate treatment is morally wrong because it is wrong to treat people differently for morally illegitimate reasons. For instance, the way an individual is racialized is not an appropriate reason for distributing benefits away from them. So, if a hiring system decreases the score of an applicant by a fixed amount on the sole basis of their racial profile, it exhibits a morally problematic procedural bias. We should note that a procedural bias (in our definition) is not necessarily morally problematic, such as when the hireability score is influenced by the class variable of education.

The process of bias reduction involves several (not necessarily sequential) phases:

(1) In the detection phase, we apply XAI methods to detect a (potentially problematic) bias. For example, we may detect that a particular feature, such as the zip code of an applicant's residential address, surprisingly influences their hireability score.

(2) In the interpretation phase, we interpret XAI's output and attribute meaning to a bias. Here, we would note that the zip code's influence on hireability score seems like a bizarre phenomenon and ask whether this indicates a bias with broader significance.

(3) In the evaluation phase, we evaluate a bias to determine whether it is problematic (morally, epistemically, or socially). This phase necessarily involves normative considerations alongside knowledge of social context.

(4) In the mitigation phase, we devise means to reduce this bias and its impact. This phase will involve not only normative but also practical considerations.

Now, we are ready to characterize TXAI, according to which:

> Technical experts (specialists in XAI methods) can handle the detection and interpretation phases satisfactorily, in relative independence from other stakeholders—that is, without extensive interaction and consultation with them.

TXAI need not deny that in the evaluation and mitigation phases, a larger group of stakeholders, along with their attendant background knowledge and interpretive resources, should be involved. In fact, in our survey of XAI literature, many researchers highlight the importance of different stakeholders. For example, Arrieta and colleagues (Arrieta et al. 2020) declare that the most important task for XAI is to carefully review the interests, demands, and requirements of all stakeholders interacting with the system. Langer and colleagues (Langer et al. 2021) argue that the success of an explanation is determined by its consideration of stakeholders' desiderata, such as their interests, needs, or expectations.

Common to these studies is the claim that XAI ought to enhance understanding of the AI system that meets different stakeholders' requirements. However, there is no mention of whether, why, and how different stakeholders can contribute to XAI for bias reduction. Given the ethical and social significance of XAI, this omission is problematic. Further, the dearth of discussion of the need to involve and integrate different stakeholders' interpretive resources for the detection and interpretation of bias indicates an attitude of "technological solutionism"—that is, the idea that every social problem has a purely technological fix (Morozov 2014).

One potential motivation behind TXAI is this: the XAI model is a relatively general-purpose and objective observational tool. That is, XAI helps us better understand everything in the AI system, only embodying minimal and objective information that facilitates the revelation of the AI decision process. Importantly, according to this view, XAI does not embody values and assumptions that might limit or distort what can be revealed. In other words, it assumes that to construct a general-purpose, objective observational tool, we only need a relatively small knowledge base: one that is, in practice, sufficiently possessed by a small group of technical experts. Broader meaning-making and knowledge-producing resources—including deep social knowledge regarding marginalized groups who are often the victims of biases and oppression—are unnecessary. Relatedly, TXAI assumes that the meaning of XAI's explanatory output is relatively transparent and requires minimal interpretive resources for its proper interpretation. Other stakeholders, such as ethicists, social scientists, policymakers, and so forth, can take XAI's output as given and move on to evaluate its moral and social significance.

In the rest of this paper, we will narrow down our focus to problematize the bias detection phase of TXAI. Let us return to the scenario discussed in the introduction: you are faced with the challenge of ensuring an opaque AI recruitment system is free from problematic biases. You reach out to an XAI team for help. After running a popular XAI method, Kernel SHapley Additive exPlanations (Kernel SHAP, hereafter SHAP), the team returns with the good news that problematic procedural biases do not play a significant role in influencing AI suggestions.

To illustrate how TXAI detects biases in this scenario, we constructed an AI recruitment system with a neural-network model. The model is trained to fully capture the decision rules embodied in the hiring data along precisely the lines of our hypothetical example. Specifically, we generated deliberately simplified hiring data, designed to emulate real-life hiring data with only the hiring decisions and five features of applicants (teamwork skills, relevant experience, willingness to work night shifts, gender, and race), instead of the usual number of around a hundred features. A total of 120 applications (instead of the usual thousands) were generated and then fed into the AI system for training and prediction before the AI system was interrogated by SHAP. Even though the neural network, hiring data, and applicant pool we constructed are simpler than those in real-life cases, this simplification does not affect the nature of bias detection. In fact, simplification makes the job of the XAI team much easier than it is in reality—being able to consider each applicant and their features in detail is unlikely *in practice* due to the larger number of feature-applicant pairs (typically in the range of themillions).

Now, let's walk through how a TXAI team typically reaches conclusions via SHAP. First, the SHAP values of the five features for each candidate are calculated. Roughly, SHAP calculates the relative importance of features (e.g., class variables) in driving the decision of a specific case in the following way: (i) It probes an AI system with different variations of the feature values of a particular candidate and observes how these variations alter the decision for the candidate (e.g., their hireability scores) (Lundberg and Lee 2017). Then, (ii) it uses a simple linear regression model to fit its observation, generating the feature-importance values based on the regression coefficient of each feature, respectively. With the feature-importance value calculated, SHAP can highlight a subset of features that play an important role in determining the scores of a particular candidate.

Because the numbers of applicants and features are often so massive in realistic situations, a team would look for group-level patterns. Figure 1 illustrates the SHAP values for our 120 applicants through a group-level "force plot" (more about this soon). It shows that ability to work night shifts and relevant experience are the

main features that contribute positively (indicated by the color red) to the decision to hire a candidate, while inability to work night shifts is the main feature that contributes to the decision to reject a candidate (indicated by the color blue). Indeed, the absolute SHAP values for each feature averaged across all cases (shown in figure 2) confirm the observation above: the most important features driving the decisions are the ability to work night shifts and relevant experience, while gender, race, and teamwork skills each appears to play only a minor role.

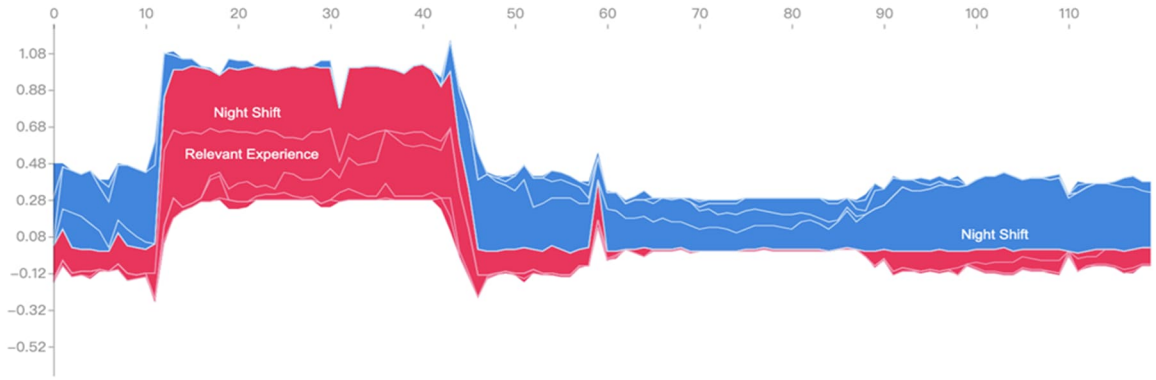

Figure 1. The group-level force plot of SHAP values for 120 candidates. It is composed of 120 horizontally stacked individual force plots (figure 3) of each candidate.

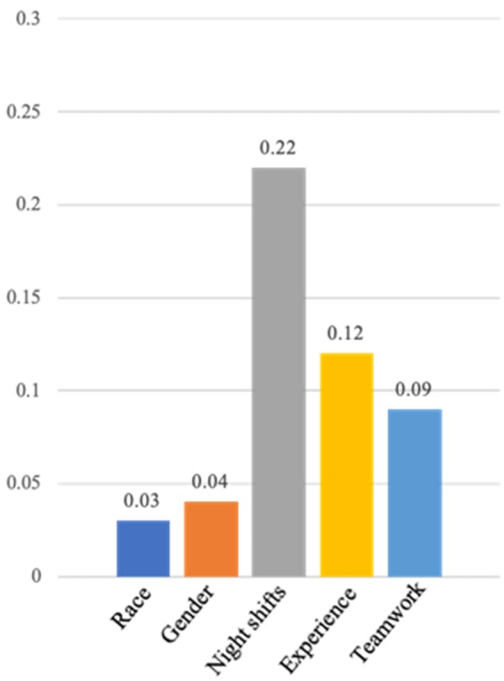

Figure 2. The absolute SHAP values for each feature (averaged across the group of 120 applicants).

To better understand how the group-level pattern is produced, we looked into the SHAP values of individual cases. Figure 3 shows the force plot of SHAP values of two candidates, visualizing the relative feature importance in corresponding hiring decisions. The upper panel illustrates the feature importance in driving the decision to hire (indicated by the score of 1.00) a particular (male) candidate. Willingness to work night shifts is the most important feature that positively contributes to the decision (as indicated by the color red), followed by his relevant experience and teamwork ability. His gender contributes positively, and his race contributes negatively (as indicated by the color blue), both in relatively minor ways. The lower panel illustrates what drives the decision to reject (indicated by the score of 0.00) a particular (female) candidate. It is primarily her inability to work night shifts and lack of relevant experiences that contribute to the rejection (as indicated by the color blue), despite her teamwork skill, gender, and race having some minor positive contributions to her hireability score (as indicated by the color red).

Figure 3. Upper panel: the individual-level force plot of SHAP values for one candidate recommended for hire by the AI system. Lower panel: the individual-level force plot of SHAP values for one candidate recommended for rejection.

Based on the observation that the group-level SHAP values of race and gender are close to zero, an XAI team would likely conclude that these features do not significantly influence overall decisions. That is, the AI system does not have significant procedural biases based on gender and race; therefore, no bias-reduction procedures need to be applied to this system.

One important qualification must be made before we end this example: one may find the decision not to intervene in this AI system objectionable. After all, gender and race are in fact influencing hiring decisions,

no matter how small the impact is. Moreover, aren't detecting and removing subtle biases like these the main motivations for XAI?

An important reason to treat very minor influences like this as mere noise is that minor influences are often introduced during machine learning processes. For example, figure 4 shows the force plot of an AI system trained on artificially generated hiring data free from procedural biases concerning gender and race, together with the absolute SHAP values for each feature averaged across all candidates. As we can see, there are still minor influences of race and gender on this AI system's decisions, which reflect statistical fluctuations during learning of the AI system and/or the testing of the XAI procedure. As a result, XAI teams will have to ignore these minor influences as noise.

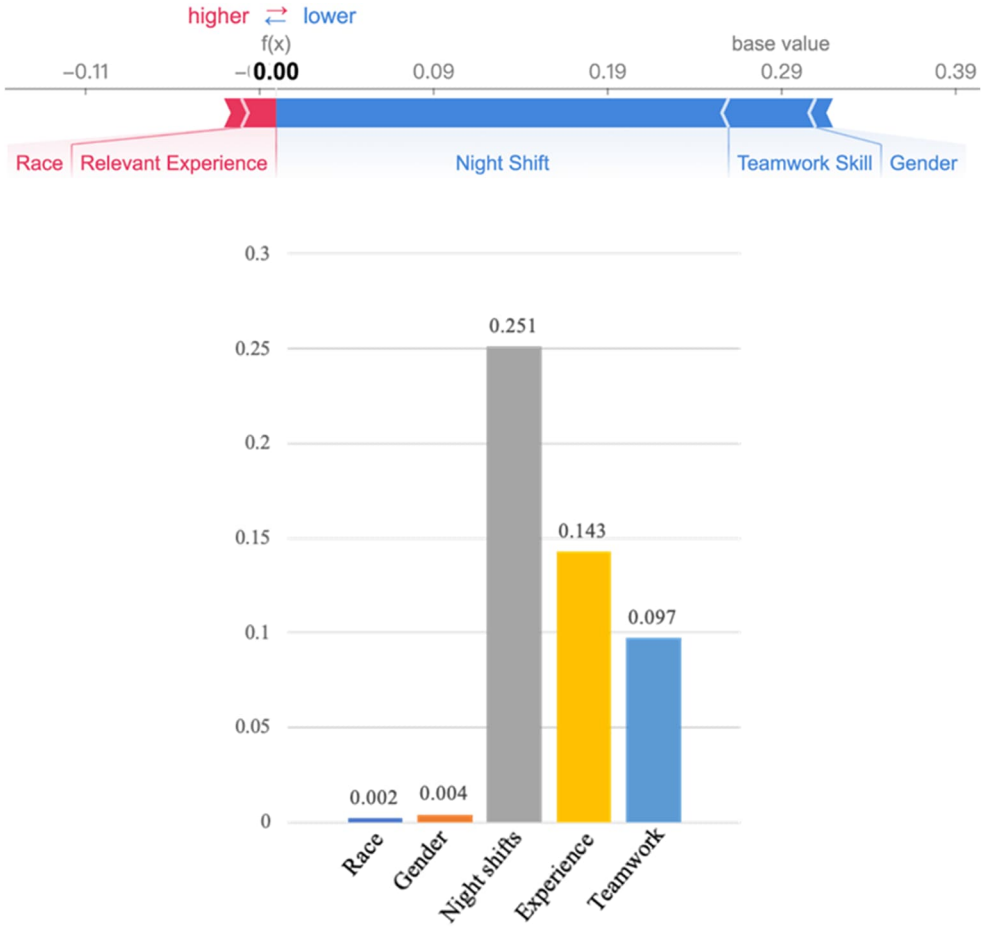

Figure 4. Upper panel: the individual-level force plot of SHAP values for one candidate recommended for rejection by an AI system trained with hiring data with no procedural biases based on gender and race. Lower panel: the absolute SHAP values for each feature (averaged across all applicants).

Even if one is not convinced by the above practical consideration, our hypothetical case study of what XAI teams would conclude about the AI system will still serve its illustrative purpose: to show that TXAI is not reliable for bias detection. As we will reveal later in section 4.2, conclusions drawn by the TXAI team are open to vigorous challenge.

Having defined TXAI and illustrated it in action, we now situate it within the ethical AI movement (Hao 2019; Kind 2020; Häußermann and Lütge 2022). The first wave of AI ethics was characterized by the quest for appropriate principles that should guide the development and use of AI systems. Then, the second wave took a technical turn: researchers strived to find technical solutions to questions such as fairness and discrimination. Now, at the threshold of an emerging third wave, it has become clear that AI systems are sociotechnical systems that "cannot be understood outside of the social context in which they are deployed, and they cannot be optimised for societally beneficial and acceptable outcomes through technical tweaks alone" (Kind 2020). In line with the third wave of the ethical AI movement, we will argue that TXAI, an instance of the technical turn, is inadequate. We will argue for the integration of XAI into a broader social-epistemic process in what follows.

3. Feminist Insights

Questions of power imbalances, social justice, and how new technologies can bring about actionable changes are now taking center stage. These topics have long been a central concern of feminist philosophy. Therefore, in considering the ethical and social implications of using XAI for bias detection, we turn to feminist theories—in particular, feminist philosophy of science and discussions of situated knowledge and epistemic injustice—for inspiration. These analyses are especially helpful, as they stress the importance of context, uncover how values permeate our epistemic inquiries, and examine the ways extant power disadvantages those in marginalized groups in the production, distribution, and justification of knowledge. Below we selectively discuss parts of these critical lenses most relevant to our current purposes.

3.1 Feminist Philosophy of Science

Feminist philosophers of science have long argued that science is deeply value-laden and typically reflects and reinforces the ways of thinking and doing of socially privileged groups (e.g., Nelson 1990; Longino 1995, 1996; Douglas 2000, 2009, 2016). Community practices, including scientific ones, are shaped by social values; whether the practitioners are conscious of these influences or not, scientific practices are replete with

value-laden, contingent decisions.[6] What is studied and how—which aspects of the phenomenon are a focus, the choice of research questions, investigative methodology, framework assumptions, interpretive strategies, standards of empirical adequacy, and so on—are by no means value-free.[7]

One important focus concerns how social values determine the assumptions scientists adopt to mediate the relationship between hypotheses and data (Longino 1995, 1996, 2001). Scientific theorizing relies on inductive methods: researchers derive patterns from existing cases and project that those patterns will hold for novel cases. Yet evidence based on past experience alone is never sufficient to prove ampliative conclusions, and assumptions are required to guide the researchers' inferences. While nonevidential assumptions are essential to inductive inference, the ideal of value-free science allows for only *epistemic* nonevidential assumptions, such as epistemic theoretical virtues (Kuhn 1977) . By contrast, *nonepistemic* assumptions—those shaped by social values—have no place in scientific practice.[8] However, Longino (1995, 1996) forcefully argues that the distinction between epistemic and nonepistemic theoretical virtues ultimately fails, as the demarcation between the two itself is the result of social values, hence undermining the possibility of the value-free ideal.[9]

Recently, Johnson (forthcoming) contends that a similar value-laden practice recurs in the domain of machine learning and argues that it is constitutively value-laden. As she argues, predictive algorithms are inductive procedures by nature, so they too are under the influence of nonevidential and nonepistemic values. For example, one of the most widely discussed issues of machine learning programs is the data problem: what the training data consists of, how data are gathered, and what data must be taken into consideration are all very important theoretical questions with serious practical consequences (D'Ignazio and Klein 2020). Specifically, training data, in practice, are usually based on certain default groups; typically, these groups are socially privileged and prioritized in theory (i.e., considered as a representative group). Generalizations built around privileged groups, however, may not be applicable to the entire population, and the resulting program may generate unrepresentative judgments that negatively impact marginalized groups in particular. For example,

[6] See Brown (2020) for discussions on how contingent decisions in science can be made in many different ways and can impact overall performance, users, and the public.

[7] See Kourany's (2010) discussion of a socially responsible science, as well as Brigandt (2015) on how values can influence a theory's conditions of adequacy.

[8] Values come in a wide variety, including spiritual, aesthetic, environmental, ethical, and political, among many others. For the purpose of this paper, we focus on the contrast between epistemic values—such as a theory's internal consistency, predictive accuracy, fit with evidence, simplicity, etc.—and those that are nonepistemic, which we will refer to as *social values* (Brigandt 2015).

[9] See also Laudan (2004), Lacey (2004), and Biddle (2020).

Joy Buolamwini, a Ghanaian-American computer scientist and digital activist, discovered that facial recognition software had trouble detecting her dark-skinned face but no difficulty recognizing her lighter-skinned collaborators.[10]

Note, however, that in this case the machine learning algorithm is not obviously guided by nonepistemic values. That is, when the algorithm takes privileged groups as the basis for making inductive inferences, it can be seen as promoting the epistemic value of simplicity. However, as Johnson (forthcoming) argues, in this context, "an allegiance to simplicity will (either wittingly or unwittingly) imbibe the very socio-political values on which the hierarchical relations are formed," hence rendering the practice constitutively value-laden.

3.2 Situated Knowledge & Epistemic Injustice

Feminist epistemology has long argued for the situatedness of knowledge.[11] Perspectival differences make epistemic differences (Grasswick 2018). What we know and how we know are intimately connected to our values and assumptions, which are determined by our material and social positions. For example, Haraway's (1988) work challenges the myth of scientific objectivity and exposes this so-called objectivity as unreflexive ignorance of one's own biases. She offers the alternative of *situated knowledge*: knowledge is produced by specific people in specific historical, geographical, and cultural circumstances. Since every single epistemic agent is socially situated, all forms of epistemic endeavors are socially situated.[12]

Once we recognize the essential situatedness of all knowing, it follows that all perspectives, including the views of the dominant and privileged, are partial. As different individuals and groups governed by distinct values and assumptions are likely to make different choices in the process of knowledge production, they form different "local epistemologies." Centering on only a small set of perspectives can easily result in distorted accounts of the whole social order (Harding 1991, 1993). By contrast, expanding the backgrounds

[10] Steve Lohr, "Facial Recognition Is Accurate, If You're a White Guy," *New York Times*, February 9, 2018, https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[11] Situated knowledge is a claim shared by feminist empiricists and standpoint theorists. For further distinctions between the two approaches, see, for example, Intemann (2010) and Anderson (2020).

[12] As Alcoff and Potter 1993 and Alcoff 2001 point out, each of us comes to knowledge-making processes from a particular combination of perspectives, so the perspectivality at work can be extremely complicated and it is difficult to determine exactly what combination of perspectives are making the difference in any given case. Further, granted that all theorizing proceeds from some epistemic location, Louise Antony (2001, 142) cautions that we must treat the truth-conduciveness of particular biases as an empirical question. Similarly, the truth-conduciveness of any particular perspective should be treated as an empirical, not a priori, question in our ongoing pursuit of knowledge.

of epistemic agents involved in the production of knowledge will be advantageous in general, since unexamined, or underexamined, features of relevance will be given due attention while challenging problematic assumptions. Therefore, we should aim to render knowledge-making processes more inclusive. Arguably, the perspectives of the minoritized can be particularly valuable not only for their potentially special access to the deep knowledge of society (Anderson 2020) but also for the opportunities they provide for researchers to reflexively examine their own social positions and privileges.

Recent work on epistemic injustice (Fricker 2007, 2016; Mason 2011; Medina 2012, 2013; Dotson 2012, 2014; and many others) echoes this attention to the situatedness of knowing. Epistemic injustice concerns how members of marginalized groups are often systematically prevented from participating fully in collective meaning-making and epistemic processes. Most notably, Fricker (2007) identifies two types of epistemic injustice: testimonial injustice occurs when an epistemic agent suffers from unwarranted prejudicial credibility deficit due to their social identity; hermeneutical injustice occurs when there is a gap in the collective interpretive repertoire, such that marginalized social groups are at an unfair disadvantage in making sense of, or communicating, their socially important experiences. In both cases, victims of epistemic injustice are unfairly diminished in their capacities as knowers.

Drawing upon this work, Pohlhaus cautions against *willful hermeneutical ignorance*: the problematic propensity to dismiss alternative epistemic resources put forth by marginalized knowers such that the dominant knowers run the risk of "continu[ing] to misunderstand and misinterpret the world" (2012, 716). Importantly, minoritized groups should not bear the unfair burden of proving their epistemic worth (Berenstain 2016), and their perspectives ought to be taken more seriously because, in practice, they are likely to offer invaluable interpretive resources, rooted in their distinctive experiences, that help fill in lacunae in collective meaning-making resources (Goetze 2018). In fact, the failure to take heed of their perspectives can be a huge loss to the entire community, as potentially useful epistemic resources are excluded from our pursuit of knowledge.

In short, feminist epistemology studies our understanding of knowledge, objectivity, and scientific methodology. It reminds us of the significance of context, as well as the influence of values and power in our epistemic activities, such as that exemplified by TXAI. We now turn to a feminist reading of TXAI and demonstrate how the lack of diverse perspectives can negatively influence the potential of XAI for bias detection.

4. Why XAI Needs Feminist Theories

Drawing on feminist insights, we argue in this section that TXAI is mistaken. We first present our conceptual argument before returning to the case study introduced in section 2. The main driving force behind our argument is the importance of context. Whether an algorithm is biased is highly *context-dependent* and requires an equally *context-sensitive* observational instrument for its detection. However, TXAI will likely not, in practice, have such a high level of context sensitivity. So, TXAI cannot be a reliable process for detecting algorithmic biases.

4.1 The Argument against Technical XAI

Our argument starts with an uncontroversial premise:

> (1) A highly context-dependent property can only be detected with rich and relevant interpretive resources.

A context-independent property requires minimal interpretive resources to identify because it has *invariable, intrinsic* defining features. For instance, being a triangle is not context-dependent. Having exactly three sides define triangle: this feature is invariable across contexts; it is also intrinsic because it does not depend on anything beyond the object.

By contrast, a highly context-dependent property involves *relational* features. For example, being a member of a particular social class is a context-dependent property and involves relational features such as occupying a certain social role—a feature one can possess only in relation to a social system. One cannot be a king without a kingdom, so to speak. Moreover, being a king is not an intrinsic property of a person—one can be dethroned and lose all political power while nothing intrinsic has changed. As a result, to detect a highly context-dependent property, we need to look beyond the entity itself into its relations with others within a broader context.

Being a member of, say, the most privileged social class also involves *variable* intrinsic and relational features, both within and across contexts. Such members look different and occupy different social roles not only between, say, ancient Egyptian and contemporary US societies, but also within the same society. That is, they form a highly heterogeneous set. As a result, there are no simple, invariant low-level features that can be used for identification.

Hence, to identify a highly context-dependent property, we must observe a wide range of intrinsic and relational features. Moreover, to properly infer from what we observe, we need to be equipped with rich interpretive resources about specific contexts.[13]

Our second premise concerns the nature of bias:

> (2) Being a particular type of bias is a highly context-dependent property; as a result, to identify it requires rich and relevant background assumptions.

First, what distinguishes between particular types of biases will involve relational features. Consider two biases represented by the same mathematical function and, as a result, intrinsically identical: one may represent how the feature of an applicant's *gender* affects the hiring decision, yet the other may represent how the applicant's *relevant experience* affects the hiring decision. This is because what a bias represents is not determined entirely by what goes on in the AI system but depends on how the system interacts with the larger world (i.e., meaning is determined by causal-historical facts of the interactions).[14] Thus, to identify a specific type of bias (say, bias based on gender versus bias based on relevant experience), one must examine its relations within a broader context.

Second, being a particular type of bias also involves highly variable intrinsic and relational features. A particular type of bias (e.g., gender bias) can be implemented in different ways in the same or different AI systems (e.g., they implement different mathematical functions, such as a linear or a curved function). So, the same type of bias will have different intrinsic features within or across contexts. There are also variable ways a particular type of bias relates to the larger context. For example, a gender bias can come to represent the influence of gender due to a wide variety of ways AI systems interact with the social world, such as the impact of gender on the burdens of child care, access to proper health care, and so on. Hence, there are no invariable intrinsic or relational features one can use to identify biases.

[13] In a similar vein, Fodor (2000) discusses the issue of context-dependent properties regarding relevance-determination and modularity. Also relevant is a point drawn from feminist empiricism that data only supports a hypothesis against background assumptions (Longino 2001).

[14] For instance, zip codes in the United States are often an indicator of race. So, if the AI system's hiring recommendation filters out a candidate based on their zip code, there is likely to be a problematic racial bias. Without knowledge or awareness of the significant housing segregation, however, one can easily overlook this. See also Cappelen and Dever (2021) for a general argument for an externalist metasemantics for AI.

In short, being a particular type of bias is a highly context-dependent property. It is more similar to being a member of a particular social class than being a triangle. They have no invariant intrinsic or relational low-level features that one can use to detect within or across contexts. As a result, detecting a particular type of bias, like identifying a member of the most privileged social class, requires a wealth of interpretive resources.

We now move to the most substantial premise of this argument:

> (3) TXAI will not, in practice, always have the rich, relevant interpretive resources necessary to flexibly design and deploy appropriate XAI models for detecting a particular type of bias.

Unpacking (3) involves a number of steps. First, we will problematize the conception of XAI models as general-purpose observational tools. Second, we will apply feminist insights to reveal that XAI models are deeply value-laden. Third, XAI models, when not properly informed by relevant interpretive resources, may fail to provide appropriate explanations given the goal of detecting a particular type of bias. Finally, the interpretive resources needed may be systematically excluded in the XAI model's design process.

XAI models are not general-purpose. XAI models are, at best, *special-purpose* observational tools—that is, they can only accurately detect a limited range of biases. This is because there is an unavoidable conflict between explainability and comprehensive accuracy: to facilitate interpretability, an XAI model is an intentionally simplified representation of a complex AI system. With such a limited representational resource, an XAI model cannot represent *all* aspects accurately. Like most models, some features of its representational target are *abstracted,* and some unrealistic features are incorporated as *idealizations* (Weisberg 2007).

For the purpose of bias detection, an XAI model should offer a (more or less) accurate representation of the selected types of biases while offering a highly idealized (thus intentionally false) or abstract (thus intentionally partial) representation of the other aspects of the AI system.[15] Importantly, what an XAI model can represent accurately will depend upon the simplifying assumptions adopted to build the model. Only when there is an appropriate attunement between the simplifying assumptions and a particular bias (whether achieved intentionally or not) can the XAI model represent the bias accurately (for a concrete illustration, see section 4.2).

[15] While most philosophers of model sciences will agree with the conception of idealization as intentional falsehood, some disagree—for example, Nguyen (2020).

XAI models are not objective. XAI models are not *objective* observational tools, as social values and assumptions have a more significant impact than is often recognized. As feminist philosophy of science points out, there are no objective models or theories in the traditional sense—a theory or model necessarily embodies social values and assumptions even under ideal conditions. In fact, the context of XAI model-building is far from this objective ideal: XAI models need to balance, on the one hand, the concerns over explainability and simplicity, and on the other hand, truth and other epistemic norms (e.g., representing AI systems accurately to reflect the complexity of information processing). As Potochnik (2012) points out, the tradeoff between different aims of modeling mirrors the discussion of the value-laden choice of theoretical virtues. Specifically, since the simplifying assumptions adopted in an XAI model already must be highly idealized (a strictly false or mismatched fit between model and the world) or abstract (omitting details), considerations other than facts and epistemic norms naturally (and sometimes appropriately) come to play a larger role at various decision points, irrespective of designer intentions.

Not all XAI models are appropriate for detecting a specific type of bias. First, different XAI models provide different explanations. This follows from applying the insight of situated knowledge to model sciences. As Potochnik (2012, 386) puts it:

> Different models will represent different features of a system, at different degrees of abstraction; they will employ different idealizations; . . . a commitment to the idea that different values generate different scientific knowledge about a single phenomenon does not involve granting the truth of multiple claims that are mutually inconsistent. Instead, it amounts to the much less problematic idea that multiple models provide different representations of the target system.

Due to different simplifying assumptions, different XAI models will provide different explanations, which latch on to different causal patterns of the same system in different ways.

However, given the goal of detecting a specific type of bias, not all explanations are equally acceptable. To provide adequate explanations, the XAI model will need to embody simplifying assumptions attuned to the type of bias it aims to detect. To this end, researchers need to be informed by relevant epistemic and hermeneutical resources, some of which may only be available in local epistemologies of marginalized groups targeted by this bias. Indeed, this is why marginalized groups, by reflecting on their unique experiences, may gain privileged access to certain perspectival knowledge that others cannot easily acquire.

We are not saying that those without marginalized backgrounds could never have the requisite interpretive resources. On the other hand, some technical experts may even come from marginalized backgrounds themselves. Still, technical experts, in practice, tend to belong to selective local epistemologies. They are likely to be influenced by similar assumptions and values, focus on certain aspects of AI systems, ask the same questions, and adopt typical methodologies. Consequently, they will often unreflectively adopt certain simplifying assumptions and practices in building XAI models, some of which may be inappropriate for detecting a particular type of bias.[16]

XAI model-building can be epistemically unjust. Importantly, these necessary meaning-making and knowledge-producing resources are often systematically prevented from entering the mainstream epistemic community. As a result, technical experts are unlikely to have ready access to them. That is, epistemic injustice—such as testimonial and hermeneutical injustices—will systematically prevent marginalized groups from participating in, and contributing fully to, collective meaning-making and epistemic processes. XAI's model-building process is precisely one such instance.

By bringing these key analyses together, we can show why TXAI is unlikely to have the necessary resources to develop appropriate XAI models: (i) XAI models, *qua* special-purpose tools, can detect only those biases that their simplifying assumptions are attuned to. (ii) These simplifying assumptions are determined in part by the social values and assumptions of technical experts. (iii) Whether an XAI model for bias detection works depends on whether technical experts can alter its simplifying assumptions via reflexive awareness of the influence of their own social values and background assumptions. (iv) However, the relevant interpretive resources are often not incorporated into mainstream epistemic resources; as a result, a small group of technical experts, in practice, are unlikely to have ready access. (v) Moreover, given the practice of TXAI, technical experts work in relative independence from diverse stakeholders, and this lack of interaction further prevents them from appropriately adjusting their choice of simplifying assumptions.

As a result,

(4) TXAI is not an epistemically reliable process for producing XAI models for bias detection.

[16] This point is in line with de Melo-Martín and Intemann (2011). Drawing from the development of the HPV vaccines, they argue that research guided by feminist principles (including (i) diversifying researchers, (ii) taking seriously the perspectives of marginalized stakeholders who are affected the most, and (iii) making visible various social positions) would be epistemically superior and socially responsible. We thank an anonymous reviewer for this reference.

A further implication of this argument is that a more reliable process would be a social-epistemic one that could incorporate relevant background knowledge and interpretive resources. We call such a process integrated XAI. In section 5, we will draw from feminist insights to suggest how to implement this process. However, before we do so, we will return to our prior case study to illustrate our conceptual point.

4.2 Case Study

Recall the recruitment scenario we introduced at the beginning of this paper. In section 2, we illustrated how a typical XAI team under the TXAI paradigm operates. Using SHAP, technical experts conclude that there are no significant procedural biases based on gender or race in making hiring recommendations. We have also hinted that this conclusion is badly mistaken.

In fact, we intentionally generated the hiring data to embody a strong bias against black women, and the AI system learned to fully capture this bias. Specifically, our AI system reflects a neutral attitude toward black men, white men, and white women, but a strong bias against black women.

Inspired by intersectionality theory—how one's various social identities combine to constrain and enable—we call this bias *intersectional bias*. In this case study, the intersectional bias is so strong that the AI system recommended rejection of *all* black female applicants (while the average rejection rate is 72 percent), despite the fact that we constructed their average qualifications to be identical to those of the other three groups. So, far from being mere noise, race and gender exert a strong influence on hiring recommendations.

But why does SHAP fail to discern such a strong bias? This is because SHAP's simplifying assumptions (i) cannot reveal the defining features of the intersectional bias, (ii) systematically underestimate its importance in driving hiring decisions, and (iii) conceal bias when it is only suffered by a social group that constitutes a small percentage of the entire group. As a result, SHAP is inappropriate for detecting intersectional biases. First, SHAP cannot reveal the defining feature of intersectional bias because it uses a linear regression model to generate the feature-importance value. The linear regression model is most appropriate when different features influence the decision (more or less) independently and linearly. However, when they deviate from this characterization (e.g., an intersectional bias where the influence of two features depends on each other, which also entails nonlinearity), the model cannot reveal the nature of the bias. This is because, roughly, information about the nonlinearity and dependency between the two constitutive features is inevitably lost

after a linear regression model is applied: one cannot determine from the feature-importance values of gender and race (represented by figure 1) the fact that it is an intersectional bias that produces these impacts.

Second, SHAP will systematically underestimate the feature importance of intersectional bias (if we take it to be roughly reflected by the sum of the values of two constitutive features), because the way such value is generated does not respect an intuitive conception of how to explain the influence of an intersectional bias.[17] For example, when we evaluate the importance of intersectional bias against black women, the intuitive and ideal measure is to directly intervene on the features the intersectional bias targets—that is, flipping an applicant's relevant features into a contrasting class (turning a black female candidate into a "non-black female" candidate, and vice versa) and observing how it impacts decisions.

However, the way SHAP performs interventions does not respect such an ideal. SHAP "intervenes" in a particular case by trying out all combinations of variations in values of features, such as gender, race, and relevant experience. Importantly, SHAP, given its current design, cannot treat "being a black woman" as a single feature to manipulate as a unit. As a result, SHAP inappropriately (i) includes in the contrasting class combinations that should not have been there, and (ii) counts the same combination (e.g., for the purpose of evaluating intersectional bias) as distinct. For example, it may intervene in the case of a white male candidate by changing the candidate's gender or race (turning him into a black man in one intervention and a white woman in another intervention). Note that because the original candidate and the two hypothetical ones are all "non-black women" the two interventions are inappropriate interventions, and in any case, they should be counted as the same (non-)intervention. Moreover, they generate no impact on the decision. Because these and other "non-interventions" with no impact are included (inappropriately or repeatedly) in the calculation, SHAP will systematically underestimate what we intuitively take to be the feature importance of intersectional bias.

Worse still, SHAP also weighs more heavily the impact of "local" interventions (i.e., the impact generated by smaller changes in values of features) when it combines the impacts of all interventions to calculate feature-importance values. Thus, it not only includes repeatedly inappropriate interventions (such as when a white male candidate is turned into a white female or black male candidate), but it also systematically weights them more heavily than appropriate ones (such as when a white male candidate is turned into a black woman),

[17] See van Fraassen (1980, chap. 5) on the pragmatics of explanation, according to which (some) explanations answer the why-question and are context-relative.

because the former is more "local." Again, this leads to the feature importance of gender and race being further underestimated.[18]

Finally, SHAP will conceal the bias suffered by the social group that constitutes a smaller percentage of the applicants. This is due to the practice of taking the mean value (of the absolute values of feature importance of a feature for all cases) to represent the relative importance of a feature at the group level (see figure 2). We should note that this practice is not part of the modeling assumption of SHAP, but it is often necessitated in practice by the sheer number of feature-applicant pairs (in the range of millions in a realistic scenario), as discussed in section 2. It is simply impossible to examine each feature-applicant pair in order to assess the group-level pattern. Consequently, a bias suffered by only a minority of applicants (such as black women, who make up the smallest percentage of applicants—10 out of 120, in the applicant pool we constructed) will be concealed. That is, the feature-importance values, even if high for those candidates, are "diluted" after averaging and will look much less significant.

One counterargument could be that TXAI has nothing to worry about until it is shown to face challenges in realistic contexts. The first concern is that our case study is hypothetical and purpose-built, and cannot be generalized to real-life situations or to other XAI methods. However, there is nothing contrived about our case.[19] Crucially, in our example, the SHAP method fails because intersectional bias (i) implements a nonlinear function, (ii) has two interdependent features (intersectionality), and (iii) impacts a minority group. All are common features of biases and easily overlooked by predominant TXAI methodologies. It also does not help to adopt another XAI method (without doing so flexibly, informed by the specificity of the bias) because it will simply make another set of simplifying assumptions that inevitably miss some other types of bias.

The second concern is that we have exaggerated the importance of interpretive resources for bias detection in realistic contexts. It may be true that it is hard to detect a specific bias if the technical experts know nothing about its nature. But once they know of the existence of particular biases, they can consequently narrow down

[18] We have simplified the explanation of the second reason by discussing it as if SHAP treats gender and race properly as categorical features, but this is not always the case. When they are treated inappropriately as continuous features, the problem of underestimating feature importance of the intersectional bias is further exaggerated.

[19] Here is an actual case of intersectional bias: *DeGraffenreid v. General Motors Assembly Div.*, Etc. (413 F. Supp. 142 [E.D. Mo. 1976]). In 1974, General Motors fired every black woman working at its assembly plant in St. Louis. DeGraffenreid sued GM, claiming that it had violated Title VII of the Civil Rights Act, which prohibits discrimination on the basis of gender and race. DeGraffenreid's case, however, was dismissed. The judge contended that since GM had (white) female and (black) male employees, it did not discriminate against women or blacks.

the search space. This would mean paying attention to only information related to, say, gender and race, and performing an exhaustive search for their problematic influences. We respond that (i) it is a general problem that many biases are not well known by those outside the affected group, and (ii) information related to gender and race (and semantic properties in general) is highly context-dependent with variant intrinsic and relational features. So, to flag such information accurately in an AI system with realistic complexity, rich interpretive resources are necessary.[20]

To sum up, we have illustrated how an XAI model's simplifying assumptions, when not attuned to a specific bias, can prevent its detection. We also showed how a lack of relevant background knowledge, interpretive resources, and (in our hypothetical case) understanding of intersectional bias can prevent technical experts from tackling this problem. Our case study highlights the main point of our conceptual argument: relevant background knowledge and interpretive resources need to be brought to bear during the construction of XAI models in order to ensure their reliability. TXAI, due to its relative insulation from marginalized stakeholders, is unlikely to possess such resources. Instead, we should replace it with a new vision: integrated XAI.

5. Toward an Integrated XAI

This is what we think XAI for bias detection ought to be:

> The design and deployment of XAI methods for bias reduction should be implemented in a socio-epistemic participatory process that actively incorporates the perspectives of diverse stakeholders, especially the epistemic and hermeneutical resources of marginalized groups.

We dub this socio-epistemic process integrated XAI (hereafter, IXAI). IXAI recognizes that to successfully detect (and interpret) a wide range of highly context-dependent biases, an equally context-sensitive process is required. Moreover, such a high level of context sensitivity can only be achieved in an inclusive *social* process designed to flexibly marshal the relevant epistemic and interpretive resources to reduce the influence of bias. A rich literature exists concerning how to engage diverse stakeholders in scientific practice;[21] here, we provide two specific pathways toward IXAI, drawing upon the key lessons from the above.

[20] This is related to the problem of bias interpretation confronting TXAI (mentioned in section 2), an interesting issue we plan to tackle next. See Hu (n.d.) and Johnson (n.d.) for issues related to interpreting information about race.
[21] For example, Kevin Elliott (2017, chap. 7) lays out four forms of engagement between stakeholders: bottom-up engagement, top-down engagement, interdisciplinary engagement, and engagement with the laws, institutions, and policies.

First, IXAI actively tackles epistemic injustice that prevents the incorporation of diverse and marginalized perspectives. Feminist theorists have proposed ways to address hermeneutical and testimonial injustice that could be of use here. For instance, to bridge the gap in collective interpretive resources, cultivating epistemic virtues is important in the following ways: (i) for technical experts and those from dominant groups, modesty, humility, and open-mindedness are necessary; (ii) for other stakeholders or marginalized groups, self-trust, activism, and advocacy are particularly relevant; and (iii) for the entire institution, epistemic democracy is needed.[22] Epistemically modest technical experts, for example, are more likely to consult different viewpoints and critically examine their own decision-making processes. Moreover, ideas regarding participatory governance—including algorithmic audits and the appeal to public reason (Kearns et al. 2018; Brown, Davidovic, and Hasan 2021; Binns 2018)—may help facilitate the engagement of more diverse stakeholders and render the explanations more adequate.[23]

Second, IXAI would deploy a multiplicity of XAI methods and models, each informed by a diversity of stakeholders' perspectives, in order to reliably detect different biases. Each method or model, with its simplifying assumptions reflecting some background assumptions and perspectives, will be limited in its applicability and have its own blind spots. Yet, these necessarily limited models, each informed by different epistemic and interpretive resources and attuned specifically to different biases, can be simultaneously deployed. Each viewpoint can serve as a foil for others, critiquing the central claims and commitments of each. This pluralism in XAI modeling exerts a corrective influence as problematic simplifying assumptions are controlled for by models with contrary simplifying assumptions. This pluralistic approach therefore provides a route to reliable detection of a wide range of biases.

Let us see how these insights would function in practice. Recall that TXAI fails to detect bias because SHAP's simplifying assumptions are not attuned to intersectional bias. Inspired by the vision of IXAI, you have decided to implement measures to alleviate epistemic injustice in the XAI model-building process by informing the XAI team of the rich history behind the development of intersectionality theory.[24] Equipped with this knowledge, the XAI team probes the AI system for intersectional biases based on gender and race. They

[22] See Fricker (2007), Medina (2013), and Anderson (2012) for further discussion.
[23] See also Hellman (2020, forthcoming).
[24] As Crenshaw (1989) articulates, in the case of *DeGraffenreid v. General Motors*, a company may claim that it values diversity and fairness by focusing on recruiting black and women talent, respectively. However, such measures can leave certain minorities completely unaccounted for. By hiring and promoting black men and white women, black women are ignored and further disadvantaged.

quickly realize that to detect such biases, they will need to modify their modeling method and practice. To begin, they minimally modify the practice of averaging. Instead of taking an average across the entire group, they look at SHAP values averaged across different subgroups with intersecting gender and racial identities. This modification fruitfully reveals that the subgroup of black women, unlike other subgroups, suffers from significant biases based upon gender and race. As figure 5 shows, SHAP values for race and gender for black women are much higher than those for the entire pool of candidates. This highly suggests that being a black woman is one of the main reasons for their rejection.[25] The XAI team concludes that the AI recruitment system should be suspended as it is very likely to harbor a strong procedural bias against black women.[26] That is a win for IXAI.

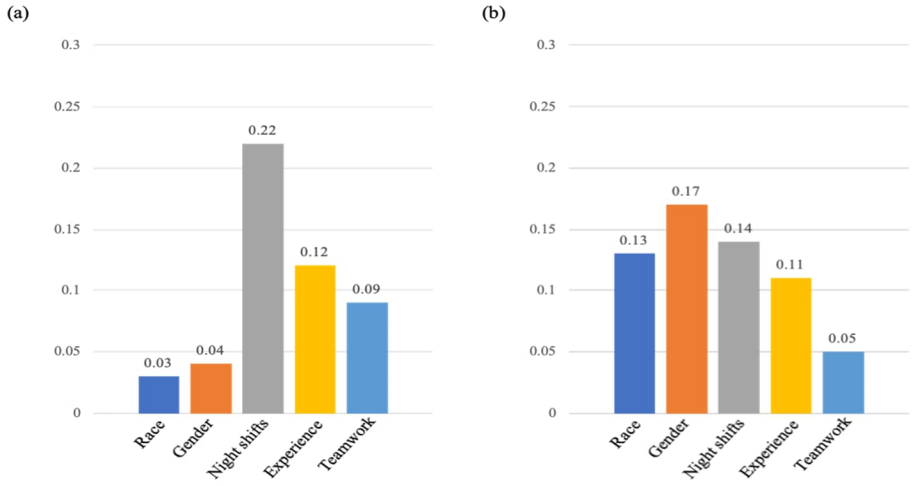

Figure 5. (a) The absolute SHAP values for each feature (averaged across the group of 120 applicants); (b) The absolute SHAP values for each feature (averaged across the subgroup of 10 black female applicants).

As Longino (2001) argues, science is a social human enterprise. We should not strive for some "objective" methodology but should actively engage diverse stakeholders in the peer-review process. This would make science truly intersubjective and about as close to "objectivity" as we can get. Our vision of IXAI, echoing insights of feminist epistemology, stresses the necessity of engaging different perspectives in XAI practice. Of

[25] However, as mentioned in section 4.2, to confirm conclusively the existence of intersectional bias against black women, an XAI method that could intervene on two features as a unit would need to be created.
[26] As a matter of fact, in 2015, Amazon decided to forgo algorithms for making hiring recommendations entirely when a similar effect was found and engineers could not correct for it reliably. See Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," Reuters, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

course, this is only the first step toward a more encompassing proposal. Our goal in this paper was to provide an impetus for the important project of shifting from *technical* XAI towards *integrated* XAI, thereby debunking the myth of a purportedly value-free methodology. The development of XAI cannot be completed by more complex technical methods or further ethical and social analyses alone. By making explainable AI a more dynamically participatory and inclusive social process, we will be in a better position to pursue more transparent, accountable, and fair algorithms.

Autobiography

Linus Ta-Lun Huang is a postdoctoral fellow at the Society of Fellows in the Humanities at the University of Hong Kong. His work focuses on philosophy of cognitive science, technology, and artificial intelligence.

Hsiang-Yun Chen is an assistant research fellow at Academia Sinica. She works primarily in philosophy of language and feminist philosophy.

Ying-Tung Lin is an associate professor at National Yang Ming Chiao Tung University. Her research focuses on philosophy of mind, philosophy of cognitive science, and neuroethics.

Tsung-Ren Huang is an associate professor at National Taiwan University. His research areas include psychoinformatics, neuroinformatics, artificial intelligence, and social robotics.

Tzu-Wei Hung is an associate research fellow at Academia Sinica. His fields of interest include the philosophy of cognitive science and the philosophy of language.

Acknowledgment

For helpful discussions and feedback on earlier drafts of this work, thanks to Po-An Chen, Josh Dever, Anson Fehross, Zhenrong Gan, Sally Haslanger, Gabbrielle Johnson, Shen-yi Liao, Chung-chieh (Ken) Shan, Mike Stuart, participants of the Feminism, Social Justice, and AI workshop (especially Carla Fehr, Karen Frost-Arnold, Trystan S. Goetze, and Catherine Stinson) and the Interdisciplinary Lunchtime Seminar at the Hong Kong Institute for the Humanities and Social Sciences (especially Herman Cappelen and Rachel Sterken), as well as the two anonymous referees. This work is funded in part by the grants 109-2410-H-001-095-MY2 (to Hsiang-Yun Chen) as well as 111-2634-F-002-004, 110-2634-F-006-022, 110-2634-F-002-045 (to Tsung-Ren Huang) from the Ministry of Science and Technology, Taiwan.

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Ahn, Yongsu, and Yu-Ru Lin. 2020. "Fairsight: Visual Analytics for Fairness in Decision Making." *IEEE Transactions on Visualization and Computer Graphics* 26, no. 1 (January): 1086–95. https://doi.org/10.1109/TVCG.2019.2934262.

Alcoff, Linda, and Elizabeth Potter. 1993. "Introduction: When Feminisms Intersect Epistemology." In *Feminist Epistemologies*, edited by Linda Alcoff and Elizabeth Potter, 1–14. New York: Routledge.

Alcoff, Linda Martín . 2001. "On Judging Epistemic Credibility: Is Social Identity Relevant?" In *Engendering Rationalities*, edited by Nancy Tuana and Sandra Morgen, 53–80. Albany: State University of New York Press.

Anderson, Elizabeth. 2012. "Epistemic Justice as a Virtue of Social Institutions." *Social Epistemology* 26 (2): 163–73. https://doi.org/10.1080/02691728.2011.652211.

———. 2020. "Feminist Epistemology and Philosophy of Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020 edition. https://plato.stanford.edu/archives/spr2020/entries/feminism-epistemology/.

Antony, Louise M. 2001. "Quine as Feminist: The Radical Import of Naturalized Epistemology." In *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, 2nd ed., edited by Louise M. Antony and Charlotte E. Witt, 110–53. Boulder, CO: Westview Press.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58 (June): 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity Press.

Berenstain, Nora. 2016. "Epistemic Exploitation." Ergo: An Open Access Journal of Philosophy 3:569-590.

Biddle, Justin B. 2020. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy.* 1–21.

Binns, Reuben. 2018. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31, no. 4 (December): 543–56. https://doi.org/10.1007/s13347-017-0263-5.

Brigandt, Ingo. 2015. "Social Values Influence the Adequacy Conditions of Scientific Theories: Beyond Inductive Risk." *Canadian Journal of Philosophy* 45, no. 3 (June): 326–56. https://doi.org/10.1080/00455091.2015.1079004.

Brown, Matthew J. 2020. *Science and Moral Imagination: A New Ideal for Values in Science*. Pittsburgh, PA: University of Pittsburgh Press.

Brown, Shea, Jovana Davidovic, and Ali Hasan. 2021. "The Algorithm Audit: Scoring the Algorithms that Score Us." *Big Data & Society* 8, no. 1 (January). https://doi.org/10.1177/2053951720983865.

Cappelen, Herman, and Josh Dever. 2021. *Making AI Intelligible: Philosophical Foundations*. Oxford: Oxford University Press.

Crenshaw, Kimberlé. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum*, vol. 1989, Article 8. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

Dargan, Shaveta, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. 2019. "A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning." *Archives of Computational Methods in Engineering*, 1-22.

Das, Arun, and Paul Rad. 2020. "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey." *arXiv* preprint, *arXiv:2006.11371.* https://doi.org/10.48550/arXiv.2006.11371.

Dunkelau, Jannik and Michael Leusche. 2019. Fairness-Aware Machine Learning An Extensive Overview. Unpublished manuscript.

de Melo-Martín, Inmaculada, and Kristen Intemann. 2011. "Feminist Resources for Biomedical Research: Lessons from the HPV Vaccines." *Hypatia* 26, no. 1 (Winter): 79–101.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: MIT Press.

Dotson, Kristie. 2012. "A Cautionary Tale: On Limiting Epistemic Oppression." *Frontiers: A Journal of Women's Studies* 33 (1): 24–47.

Dotson, Kristie. 2014. "Conceptualizing Epistemic Oppression." *Social Epistemology* 28 (2): 115–38. https://doi.org/10.1080/02691728.2013.782585.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67, no. 4 (December): 559–79. https://doi.org/10.1086/392855.

———. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

———. 2016. “Values in Science.” In *Oxford Handbook in the Philosophy of Science*, edited by Paul Humphreys, 609–30. Oxford: Oxford University Press.

Elliott, Kevin C. 2017. *A Tapestry of Values: An Introduction to Values in Science*. New York: Oxford University Press.

Fazelpour, Sina, and David Danks. 2021. “Algorithmic Bias: Senses, Sources, Solutions.” *Philosophy Compass* 16, no. 8 (August): e12760. https://doi.org/10.1111/phc3.12760.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

———. 2016. “Epistemic Injustice and the Preservation of Ignorance.” In *The Epistemic Dimensions of Ignorance*, edited by Rik Peels and Martijn Blaauw, 160–77. New York: Cambridge University Press.

Fodor, Jerry A. 2000. *The Mind Doesn’t Work that Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Glymour, Bruce, and Jonathan Herington. 2019. “Measuring the Biases that Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms.” In FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency. 269–278. https://doi.org/10.1145/3287560.3287573.

Goetze, Trystan S. 2018. “Hermeneutical Dissent and the Species of Hermeneutical Injustice.” *Hypatia* 33, no. 1 (Winter): 73–90. https://doi.org/10.1111/hypa.12384.

Grasswick, Heidi. 2018. “Feminist Social Epistemology.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018 edition. https://plato.stanford.edu/archives/fall2018/entries/feminist-social-epistemology/.

Gunning, David, and David W. Aha. 2019. “DARPA’s Explainable Artificial Intelligence (XAI) Program.” *AI Magazine* 40, no. 2 (Summer): 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Hao, Karen. 2019. “In 2020, Let’s Stop AI Ethics-Washing and Actually Do Something.” *MIT Technology Review,* December 27, 2019. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/.

Haraway, Donna. 1988. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” *Feminist Studies* 14, no. 3 (Fall): 575–99.

Harding, Sandra. 1991. *Whose Science? Whose Knowledge? Thinking from Women’s Lives*. Ithaca, NY: Cornell University Press.

———. 1993. “Rethinking Standpoint Epistemology: ‘What is Strong Objectivity?’” In *Feminist Epistemologies*, edited by Linda Alcoff and Elizabeth Potter, 49–82. New York: Routledge.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.

Häußermann, Johann Jakob, and Christoph Lütge. 2022. “Community-in-the-Loop: Towards Pluralistic Value Creation in AI, or—Why AI Needs Business Ethics.” *AI and Ethics* 2, no. 2 (May): 341–62. https://doi.org/10.1007/s43681-021-00047-2.

Hellström, Thomas, Virginia Dignum, and Suna Bensch. 2020. “Bias in Machine Learning: What Is It Good For?” In *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI),* co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), edited by Alessandro Saffiotti, Luciano Serafini, and Paul Lukowicz, 3–10. CEUR Workshop Proceedings, vol. 2659.

Hu, Lily. n.d. “What Is ‘Race’ in Algorithmic Discrimination on the Basis of Race?” Unpublished manuscript.

Intemann, Kristen. 2010. “25 Years of Feminist Empiricism and Standpoint Theory: Where Are We Now?” *Hypatia* 25, no. 4 (Summer): 778–96. https://doi.org/10.1111/j.1527-2001.2010.01138.x,

Johnson, Gabbrielle M. Forthcoming. “Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning.” *Journal of Moral Philosophy*, special issue on “Justice, Power, and the Ethics of Algorithmic Decision-Making.”

———. n.d. “Proxies Aren’t Intentional, They’re Intentional.” Unpublished manuscript.

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness.” In *Proceedings of the 35th International Conference on Machine Learning*, *Volume 80 of Proceedings of Machine Learning Research*, edited by Jennifer Dy and Andreas Krause, 2564–72. PMLR.

Kind, Carly. 2020. “The Term ‘Ethical AI’ Is Finally Starting to Mean Something.” *VentureBeat*, August 23, 2020. https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/.

Kourany, Janet A. 2010. Philosophy of Science After Feminism.Oxford: Oxford University Press.

Kuhn, Thomas S. 1977. “Objectivity, Value Judgement, and Theory Choice.” In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, 320–39. Chicago, IL: University of Chicago Press.

Lacey, Hugh. 2004. "Is There a Significant Distinction between Cognitive and Social Values?" In *Science, Values, and Objectivity*, edited by Peter Machamer and Gereon Wolters, 24–51. Pittsburgh, PA: University of Pittsburgh Press.

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. "What Do We Want from Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research." *Artificial Intelligence* 296 (July), 103473.

Laudan, Larry. 2004. "The Epistemic, the Cognitive, and the Social." In *Science, Values, and Objectivity*, edited by Peter Machamer and Gereon Wolters, 14–23. Pittsburgh, PA: University of Pittsburgh Press.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." *Queue* 16, no. 3 (May-June): 31–57. https://doi.org/10.1145/3236386.3241340.

Longino, Helen E. 1995. "Gender, Politics, and the Theoretical Virtues." *Synthese* 104, no. 3 (September): 383–97. https://doi.org/10.1007/BF01064506.

———. 1996. "Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy." In *Feminism, Science, and the Philosophy of Science*, edited by Lynn Hankinson Nelson and Jack Nelson, 39–58. Dordrecht, The Netherlands: Kluwer Academic Publishers.

———. 2001. The Fate of Knowledge. Princeton: Princeton University Press.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, edited by Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus, 4768–77. Red Hook, NY: Curran Associates.

Mason, Rebecca. 2011. "Two Kinds of Unknowing." *Hypatia* 26, no. 2 (Spring): 294–307. https://doi.org/10.1111/j.1527-2001.2011.01175.x.

Medina, José. 2012. "Hermeneutical Injustice and Polyphonic Contextualism: Social Silences and Shared Hermeneutical Responsibilities." *Social Epistemology* 26 (2): 201–20. https://doi.org/10.1080/02691728.2011.652214.

———. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford: Oxford University Press.

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." ACM Transactions on Interactive Intelligent Systems 11 (3–4): 24:1-24:45. https://doi.org/10.1145/3387166. Morozov, Evgeny. 2014. *To Save Everything, Click Here: The Folly of Technological Solutionism.* Reprint edition. New York: PublicAffairs.

Nelson, Lynn Hankinson. 1990, Who Knows: From Quine to a Feminist Empiricism, Philadelphia: Temple University Press.

Nguyen, James. 2020. "It's Not a Game: Accurate Representation with Toy Models." *British Journal for the Philosophy of Science* 71, no. 3 (September): 1013–41.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Nunes, Ingrid, and Dietmar Jannach. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems." *User Modeling and User-Adapted Interaction* 27, no. 3–5 (December): 393–444. https://doi.org/10.1007/s11257-017-9195-0.

O'Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group.

Pohlhaus, Gaile, Jr. 2012. "Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance." *Hypatia* 27, no. 4 (Fall): 715–35. https://doi.org/10.1111/j.1527-2001.2011.01222.x.

Potochnik, Angela. 2012. "Feminist Implications of Model-based Science." *Studies in History and Philosophy of Science Part A* 43, no. 2 (June): 383–89. https://doi.org/10.1016/j.shpsa.2011.12.033.

van Fraassen, Bas. 1980. The Scientific Image. Oxford: Oxford University Press.

Weisberg, Michael. 2007. "Three Kinds of Idealization." *Journal of Philosophy* 104, no. 12 (December): 639–59.