Rethinking Bias and Fairness in AI Through the Lens of Gender Studies

Gabriele Nino[1,*] Francesca Alessandra Lisi[2]

[1] University of Bari Aldo Moro, DIRIUM Dept., Piazza Umberto I, Bari, 70121, Italy
[2] University of Bari Aldo Moro, DiB Dept. & CISCuG, via E. Orabona 4, Bari, 70125, Italy

Abstract

This paper examines the main approaches that have been put forth to contrast the emergence of biases in AI systems, namely causal, counterfactual reasoning, and constructivist methodology. The objective is to demonstrate the necessity of supplementing this technical solution with a more comprehensive social analysis of the genesis of discriminatory practices. To investigate this sphere, we leverage results from the field of Gender Studies. In particular, we apply the theory of gender performativity as theorized by Judith Butler. This illustrates how AI functions within the social fabric, manifesting patriarchal configurations of gender through an analysis of the notorious case of the COMPAS system for predictive justice. This approach enables an expansion of the interpretation of the concept of fairness, thereby reflecting the complex dynamics of gender production. In conclusion, the gender dimension needs to be reconsidered not as an individual feature but as a performative process. Moreover, it enables the identification of pivotal issues that must be addressed during the design, development, testing, and evaluation phases of AI systems.

Keywords

Gender Bias, Performative Theory, AI Ethics, Algorithmic Fairness, Gender Studies

1. Introduction

Over the past decades, Machine Learning (ML) technologies have spread at great speed, helping to bring about a process of transformation and change in the social fabric. ML is a major area of the Artificial Intelligence (AI) field of study which concerns the design and the development of algorithms that can improve their performance on specific tasks (especially classification and prediction) once trained on large amounts of data [1].

The interest in ML stems mainly from the fact that they are the first form of technology ever developed that has a large reserve of agency and autonomy [2]. Consider, for example, the famous case of Chat GPT, developed by Open-AI, a generative AI system capable of performing certain tasks by understanding natural language [3]. But that is not all. The vast computational power of ML allows it to analyze large amounts of data, surpassing human cognitive abilities in terms of accuracy, speed and processing power [4]. As a result, ML is opening a wide range of applications, from medicine to architecture, from engineering to finance, and so on.

Nevertheless, the outcomes yielded by these algorithms are not always equitable or transparent. In some instances, they may even serve to exacerbate the existing prevalence of discriminatory practices within the social fabric. The potential for software to malfunction, to be biased, and thus to perpetuate social discrimination is currently the subject of rigorous critical examination by some institutional bodies [5], [6], [7], [8]. Indeed, as of 2018, the European Union Agency for Fundamental Rights (FRA) has published a series of reports examining the various forms of discrimination perpetuated by these technologies. In the survey published in 2020, we find the following: "Discrimination is a crucial topic when it comes to the use of AI, because the very purpose of machine learning algorithms is to categorise, classify and separate" [8, p. 68]. This aspect is of paramount

3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE), co-located with AIxIA 2024, November 25-28, 2024, Bolzano, Italy
* Corresponding author.
gabriele.nino@uniba.it (G. Nino); francescaalessandra.lisi@uniba.it (F. A. Lisi)
0009-0008-9896-9333 (G. Nino); 0000-0001-5414-5844 (F. A. Lisi)

importance as it links the issue of discrimination to "pattern recognition," which is the process of identifying clusters that represent the behavioral patterns of individuals by modeling their preferences. But why is it possible for forms of discrimination to lurk behind this process?

The relationship between the construction of identity and the social value attached to it has been a long-standing topic of inquiry within the fields of Feminist Theory and Gender Studies [9]. Indeed, it has been observed that individuals who are identified as women, people of color, homosexuals, or indigenous people are situated within a symbolic order that ascribes a negative value to them [10], [11]. To be different from the universal subject of Cartesian inspiration is to be "worth less" [12], [13]. For example, women have historically been regarded as non-rational beings, as a consequence of which they have been perceived as too sensitive, as being slanted by the oblative spirit that defines them [14]. This has resulted in the conclusion that they are not worthy of access to the universality of male reason. As Simone de Beauvoir posits: "Woman is seen as different from man, not man as different from woman: She represents the inessential in relation to the essential. He is the Subject, the Absolute; she is the Other." [15, p. 6]. In this sense, the American philosopher Judith Butler has highlighted that in the patriarchal view is produced an essential link of mutual exclusion, whereby Women is the mimetic correlate of the universal signifier Man [16, p. 10]. This dichotomous logic also results in the exclusion from the symbolic order of all forms of subjectivity that do not conform to gender binarism, including non-binary, trans*, and intersex subjectivities. In fact, according to the Italian philosopher Chiara Bottici, all these life forms can fall into the category of "second sexes" because "in comparison to cismen, women, two-spirited, third gender, and LGBTQI+ folks […] in the current predicament are all excluded from the category of "first sex," and that they are thus mainly the object rather than the perpetrators of gender violence" [17, p. 275]. This is why identification practices can be dangerous and expose individuals to the risk of experiencing forms of discrimination and surveillance. If the social sphere is hierarchy structured, identification processes can lead to the marginalization or exclusion of certain individuals based on their relations with a particular group.

In the field of AI, this issue is addressed through the study and modeling of *algorithmic fairness* [18], [19]. It is widely acknowledged that there are various methods to guarantee the impartiality of software and to avert its potential for discriminatory outcomes. However, our objective is to draw upon a range of concepts derived from the field of Gender Studies to enhance the discourse on fairness with a sociological and philosophical lens. In particular, we wish to put forth a reinterpretation of the *theory of gender performativity*, as developed by Judith Butler in 1990 [16]. She demonstrated how the categories we typically utilize to identify individuals, particularly those pertaining to gender and race, are not merely conceptual representations of reality; rather, they are instruments through which power relations are articulated [16]. In other words, they are tools that shape our actions. Given the growing autonomy and capacity for action of ML algorithms, it is worth considering whether they might also be capable of conveying social norms that influence human behavior.

The objective is to present a more intricate definition of the gender dimension in the context of the discussion on algorithmic fairness in AI/ML. As will be demonstrated in the ensuing discussion, it cannot be reduced to a simple individual attribute [20]; rather, it encompasses a multiplicity of references and domains that must be deployed and taken into account.

Discrimination processes are therefore highly complex, intersecting multiple vectors [21]. Gender discrimination, racial discrimination, class discrimination, forms of ableism, ageism, and so on do not run parallel; rather, they reinforce each other dramatically. A feminist approach from Gender Studies to algorithmic fairness requires unpacking this level of complexity and using the computational power of ML algorithms to change the social structures that produce these forms of oppression [22], [23]. It is imperative to integrate the concept of *intersectionality* into the concept of algorithmic fairness [24]. This entails recognizing and addressing the various forms of social oppression that undermine the democratic structure of our societies. Sexism, racism, homophobia, and xenophobia are complex social processes that reinforce one another. The report commissioned by European Digital Rights (EDRi) in 2021 demonstrated how an excessive focus on the problem of debiasing in AI and ML systems has resulted in an "algorithm-centred view" that obscures the

political scope behind the development of these software [25, p. 50]. In this sense, the authors call for an intersectional approach that can enable us to address the complexity of the phenomena of discrimination. On the contrary, it is not explicated the way in which we can theorize and enact the concept of intersectionality to the case of ML. However, it is true that intersectionality represents the most fundamental lens through which to examine the diverse ways in which these processes of discrimination and oppression converge. To this end, we believe that it is necessary to demonstrate that ML technologies operate in a performative way according to the gender performativity theory developed by Butler. In this sense it is possible to describe how ML systems operate a process of signification of the body inscribing in it a specific normative criterion. To summarize then, the objective of this study is to introduce an intersectional perspective to biases in ML by reformulating the gender performativity theory developed by Butler.

1.1. Structure of the Paper

First of all, we present a comprehensive framework for elucidating the ways in which ML can potentially give rise to forms of bias and discrimination. In Section 2, we undertake a detailed examination of the ML loop, demonstrating how the learning process can be adversely affected by the presence of a multitude of forms of bias and spurious correlations. This, in turn, gives rise to a discussion of the inherent opacity of many models.

In response to the opacity of ML systems, a set of ethical and normative concepts has been developed under the name of Explainable AI (XAI). Fairness, Accountability and Transparency are the main requirements that need to be met in order to establish the ethicality of an AI system. In particular, in Section 3 we analyze with a qualitative and historical approach the concept of fairness and the main ways in which it has been formalized at a technical level. We look at causal, counterfactual reasoning and genealogical argument and show the limitations of these approaches in relation to an adequate understanding of the social dimension of gender.

In Section 4, we leverage the method of gender performativity developed by Butler, applying it to the case of AI. This is done to integrate the social dimension within the sphere of fairness. At the conclusion of this analysis, we examine the COMPAS case, a software program utilized for calculating the probability of recidivism of a defendant, as an illustrative example of this method.

Finally, we demonstrate the need to redefine the ethical principles that guide the development and evaluation of AI in order to properly understand the multiplicity of ways in which gender or racial dimensions are expressed.

2. Ethical Problems in AI/ML

In order to axiomatize the question of the propagation of forms of discrimination in ML, it is first necessary to understand its inner workings. As mentioned earlier, ML refers to computer programs that are able to improve their performance to perform a given set of tasks as optimally as possible (Mitchell, 1997). But what does it mean in concrete terms for software to be able to learn to perform tasks and to improve its own performance?

First, the ML process can be represented as a loop [18]. The first stage of the process is to measure and collect data relevant to understanding a phenomenon. This data is used to train a model to extract patterns and generalities. After the training phase, a program, if developed correctly, can make predictions about the chosen phenomenon. Often a system can also record external feedback to improve its performance.

We are thus faced with a circular process: a social phenomenon is isolated and made the object of measurement in order to train a model capable of making useful predictions and interacting with the phenomenon in question. So it comes evident that ML operates at two levels: the technical and the social [26]. Therein lies its agency [27]. Nevertheless, it is important not to consider these two levels as ontologically distinct. As we show more in detail in Section 4.2, starting from the ’90s a conspicuous stream of works has demonstrated the impossibility of disconnecting these levels [28], [29], [30], [31], [32]. Lucy Suchman in particular describes how every technological intervention

produces a new configuration and organization of social life. For this reason, it is possible to understand the continuity and the homogeneity between the technical, the material, the social organization of the world, and their constitutive *entanglement* [33]. In this sense, Orlikowski affirms that "the social and the material are considered to be inextricably related - there is no social that is not also material, and no material that is not also social" [32, p. 1437].

Sexism, racism and sexual discrimination against LGBT+ communities are the main forms of social inequality that affect and slow down the democratization processes of liberal societies [34]. So, if the social sphere is made up of a series of unequal and discriminatory relationships between individuals, we must also look at the *technological apparatus* that can reinforce and spread these particular configurations of material life. In this sense, if ML algorithms are not designed, implemented and tested properly, they can exacerbate these differences, rather than mitigate or even eliminate them.

The philosopher Bernard Stiegler uses the Platonic term *pharmakon* to describe this dual mechanism [35]. In ancient Greek, *pharmakon* means both poison and remedy [36]. More generally, technology can be a means of promoting social justice or of destroying human bonds, as in the case of the Holocaust or the atomic bomb. So how does AI fit into this dynamic?

Each of the stages in the ML lifecycle can be affected by measurement or modelling errors that lead to the production of inaccurate results. These are commonly referred to as biases. Ferrara [37] isolates at least four main forms of bias that can affect the proper functioning of ML, each of which is located at a precise stage in the loop he has illustrated above. Let's see what they are:

Representation bias is when a dataset does not correctly represent the social set of individuals.

Sampling bias is when the training data does not consider the diversity of the population.

Measurement bias is when the data collection systematically over- or under-represents certain groups.

Algorithmic bias results from the design of an algorithm that may prioritize certain attributes leading to unfair outcomes.

This brief analysis shows that the data collection and processing phases seem to be the most critical. In fact, it is well known how ML programs have extracted spurious forms of correlations between certain features from big data sets [38]. One of the most striking cases in this sense is Amazon Recruitment Software, which was developed by the American big tech company to automate and improve the efficiency of recruitment. The software, developed in 2015, was trained on the CVs of employees hired by the company over the previous decade. As the IT sector was then, and still is, male-dominated, the system gave lower scores to female candidates, creating a false correlation between a person's gender and their technical competence. While the company tried to mitigate the problem, it decided to stop using the tool in 2017 [39].

But why is the problem of spurious correlation and its unfair results so difficult to eliminate that a multi-billion-dollar company like Amazon is forced to dismiss its program?

This brings us to the second major ethical issue raised by ML. Very often, ML algorithms are defined as opaque, meaning that it is impossible to follow their inner workings. The behavior of the software can only be judged and evaluated by its results [40].

In the '60s the Argentine-Canadian philosopher Mario Bunge has isolated this problem under the concept of *black-box*. He says: «The constitution and structure of the box are altogether irrelevant to the approach under consideration, which is purely external or phenomenological. In other words, only the behavior of the system will be accounted for» [41, p. 346].

Finally, Wendy Chun has demonstrated that ML technologies can be defined as homophilic. According with the Greek roots of the term, which means love of the same, the author reveals how statistical inference performed by ML tends to reproduce the same, what has always been [42]. In this light, one of the dangers of ML is its possible conservativeness meaning that it can predict future phenomena only replicating the past.

All these elements show that ML is not a neutral technology, but that it sometimes has a negative impact on people's lives. This is why, in recent decades, the parallel proliferation of AI systems has been accompanied by the drafting of countless ethical and legal documents aimed at regulating and

controlling this field [43]. In the following section, we analyze one of the main ethical approaches that have been developed to address the problem of opacity in systems and to provide possible oversight on ML's work.

3. A Methodology for Analyzing the Algorithmic Fairness in AI

The methodology employed to conduct the subsequent analyses reported in this article is exclusively historical and qualitative in nature. In other words, we have considered the concept of algorithmic fairness as a means of addressing the issue of discrimination in AI-based decision making. We begin by tracing the history of this concept, that did not originate in the field of computer science. From a qualitative perspective, the most significant techniques through which the concept of fairness has been formalized are analyzed. In this regard, causal and counterfactual reasoning have been identified as the two most prevalent methods for evaluating the fairness of an algorithm, notably in ML. Our qualitative inquiry, however, is aimed at investigating how the gender dimension is treated in these formalizations. In this way, we were able to identify a constitutive difference in the way the gender dimension is treated in the technical computing field compared to the field of gender studies. For this reason, we attempt to develop an interdisciplinary and a comparative approach that can enable us to connect these two areas of research.

3.1. Historical Aspects

In 2018, the Association for Computing Machinery (ACM) proposed the use of three principles that must be followed to counter the opacity of systems: Fairness, Accountability and Transparency [19]. These principles have become hegemonic in the field of so-called XAI [44], [45], [46], [47]. A detailed analysis of each of these principles is beyond the scope of this study. Instead, we are interested in focusing our analysis on the notion of fairness.

The concept of fairness has a well-established history [48]. It is a concept that originated in political philosophy, within classical liberal theories, and was brought into vogue by John Rawls's important 1985 work *Justice as Fairness* [49]. Since the 1990s, the concept of fairness has also been used in sociology and economics [50]. It is only in recent times that algorithmic fairness has also begun to be discussed. However, there is fundamental disagreement about its definition in the algorithmic field.

To overcome this problem, the notion of fairness is generally conceived in terms of a descriptive phenomenological state. That is, an algorithm is said to be 'fair' if it can be said to produce no forms of discrimination and to promote equity between subjectivities [51].

We will now analyze the main formalizations at the technical level that have been produced to assess the fairness of an algorithm. In particular, we will consider how the notion of gender is treated in these perspectives to highlight the aporias and limitations, and to show the need to graft a feminist reasoning within the discussion on fairness and, more generally, on AI ethics.

3.2. Causal and Counterfactual Reasoning

The Israeli American computer scientist Judea Pearl was the first to formalize the importance of causal reasoning in ML to overcome the danger of spurious correlations. Through his famous works, he expressed the need and necessity to move away from "reasoning by association" to "causal reasoning" [52]. So, what does causal reasoning consist of and how does it interact with the sphere of gender? To answer this question, let us look at the famous case of the admission rates at the University of Berkeley in 1973 [53].

In the early 1970s, UC Berkeley's graduate school faced scrutiny when it was observed that 44% of male applicants were admitted compared to only 35% of female applicants. This apparent disparity suggested a systemic bias against female applicants. However, when admissions data were disaggregated by department, a different pattern emerged. In statistics this effect is called Simpson's paradox [54].

Simpson's Paradox occurs when a trend apparent in aggregated data reverses when the data are divided into groups. In the Berkeley case, most departments actually had higher admission rates for female than for male applicants. The aggregate data misrepresented the situation due to differences in application patterns: more women applied to highly competitive departments with lower admission rates, while men applied to less competitive departments with higher admission rates.

Judea Pearl has extensively discussed the Berkeley case in his works. He uses the case to highlight the pitfalls of misinterpreting statistical data without considering underlying causal relationships. He writes: «Department after department, the admissions decisions were consistently more favorable to women than to men» [55, p. 311]. Thus, according to him, to properly understand the phenomena of discrimination and social order, it is necessary to consider the causal relationship between the various features of which it is composed. In this case, the concept of gender must be placed in a directed acyclic graph (DAG) in order to calculate precisely the statistical relationship between three nodes: gender, choice of department and admission [55, p. 312].

In this sense causality can be seen as a method to detect discrimination and assure fairness. Causal reasoning can identify whether disparities in algorithmic outcomes are due to discriminatory practices or other factors. For example, by using causal diagrams, researchers can determine if a protected attribute (e.g., race, gender) directly influences the decision outcome or if the influence is mediated by other variables [46].

However, Pearl's causal reasoning is divided into three important stages: The first is called association, where the statistical inference is given by the relationship between gender and admission. The second is intervention, where the relationship is modulated in a DAG between gender, departmental choice and admission. Finally, Pearl argues for the need to develop a counterfactual approach based on the question “What if I had acted differently?” which could be translated as “What if men applied to more competitive departments?” [56].

Counterfactual reasoning therefore makes it possible to break through the opacity of a program and better understand the possible discriminatory dynamics underlying it, thus fulfilling the requirement of fairness.

3.3. Limitations of these Approaches

Despite the importance of this pioneering work, especially from the fields of philosophy and law, some criticism has been levelled at this theoretical framework.

First, Ziosi et al. showed how, in XAI, the notion of fairness is measured only *a posteriori*, i.e. on the basis of a program's performance. However, as we have already anticipated, AI systems are socio-technical systems, i.e. a set of practices that cannot be reduced to the technical aspect alone. The authors show the need to adopt a genealogical approach, i.e. an *a priori* perspective that can show how the social and the technical are linked, and to take into account the different forms of discrimination that a system might produce [57].

What Ziosi et al.'s genealogical reflection does not reveal, however, is that the phenomena of discrimination, which include, for example, gender and race dimensions, are multifaceted. For this reason, Hu and Kohler-Hausmann have shown how the discussion of fairness fails to consider in advance the social ontology, i.e. the dimension in which forms of discrimination are constructed [58]. In fact, the authors show how causal and counterfactual reasoning treats the gender dimension as a separate thing that exists on its own. This means that gender is only considered as an individual characteristic, and discrimination affects the group of people who share the same characteristic. In addition, Bjerring and Busch have shown how the gender dimension is thus statistically reduced to a discrete attribute possessed by a “statistical individuality” [20].

Therefore, there is an urgent need to combine the genealogical aspect of discrimination studies with a reflection on the social ontology that enables its development and dissemination. We propose below to adopt the performative approach developed by Judith Butler and Karen Barad to explore the social ontology and understand how discrimination in ML is constructed around gender [59].

4. Research Findings from Gender Studies

4.1. Gender Performative Theory

It has been demonstrated that the eradication of discrimination cannot be achieved solely through a technical approach, rather a comprehensive strategy necessitates the integration of social perspectives, complementing the technical analysis. In particular, there is a notable divergence in the conceptualization of the gender dimension between Gender Studies and AI ethics. As has been demonstrated, gender is viewed as a "sensitive feature," that is, an individual attribute. Feminist theory posits that we examine the social ontology in which discriminatory processes are produced. The concept of gender is not merely an indication of a biological distinction between individuals (for example, between men and women). Rather, it serves as an indicator of the manner in which social relations between diverse subjectivities are organized (for instance, between women and men, heterosexuals and homosexuals). The nature of these relationships is significantly influenced by the existence of power relations that result in the marginalization or exclusion of specific subjectivities. The work of American philosopher Judith Butler is oriented in this direction.

In *Gender Trouble* [16], the author rejects the hypothesis developed by Gayle Rubin that sex is a natural, biological expression of the human, and gender is a cultural representation of the former [60]. There is no *mimetic continuity* between sex and gender. Sexual categories, i.e. male and female, are not representations of a pre-existing reality. They are the result of social production. In this sense she writes “Gender ought not to be conceived merely as the cultural inscription of meaning on a pregiven sex (a juridical conception); gender must also designate the very apparatus of production whereby the sexes themselves are established. As a result, gender is not to culture as sex is to nature; gender is also the discursive/cultural means by which *sexed nature* or *a natural sex* is produced and established as *prediscursive*, prior to culture, a politically neutral surface on which culture acts” [16, p. 11].

This ordinary conception is in fact based on the ordinary view of metaphysics, which assumes that there is a substance which can be predicated of various attributes, but which essentially pre-exists the action of attribution. Gender markers are thus not the result of the process of attributing an independent reality, the biological body, but the process by which bodies inscribe themselves within a process of signification. The gender dimension, understood in this way, is a matrix of intelligibility that allows the body to be read according to certain social norms. The author writes: “In this sense, gender is not a noun, but neither is it a set of freefloating attributes, for we have seen that the substantive effect of gender is performatively produced and compelled by the regulatory practices of gender coherence. Hence, within the inherited discourse of the metaphysics of substance, gender proves to be performative— that is, constituting the identity it is purported to be. In this sense, gender is always a doing, though not a doing by a subject who might be said to preexist the deed” [16, p. 33].

Thus, we cannot conceive gender as an individual attribute of human beings, but rather as a performative process implemented through the iterative and citational action of certain social norms. This is the social ontology we need to start from if we really want to understand how gender discrimination is produced and spread in society, and if we really want to develop an effective discourse on AI ethics.

4.2. Performativity in Science and Technology Studies (STS)

Nevertheless, Butler did not directly address the subject of technology. Instead, performativity theory has been extensively utilized and examined within the domain of Science and Technology Studies (STS) [61], [62]. The primary objectives of this field are inherent in the examination of the epistemological premises that underpin diverse scientific methodologies and practices and the ethical and political ramifications that are produced in the social domain as a result [63].

In particular, Bruno Latour showed how scientific activity does not operate as a process of representing nature but is also a way of constructing scientific phenomena. Latour's polemical target

is the modern construction of the phenomenon [64]. From a Kantian perspective, for example, the phenomenon is that which manifests itself to the transcendental subject through its categorical apparatus. In fact, modern perspective is grounded on two different transcendental poles: the objective and the subjective. The division of the world into two poles establishes a series of other dichotomies: nature/culture, necessity/freedom, immanence/transcendence, science/politics and finally non-human/human. Instead, he shows that a scientific fact never pre-exists its construction. Through the concept of the network, he breaks through the ontological barrier that opposes the subjective to the objective. The network is precisely the complex assemblage of natural forces, technical apparatuses, human and non-human agents that are connected in a more or less stable way [29]. So scientific practices do not represent anything, they produce new connections and new hybrids. It is only when a network stabilizes, i.e. becomes permanent, that the processes of signification of its constituent elements can be traced.

Latour's commitment to denouncing the simplicity of the modern stance in scientific practice has been taken up and reinterpreted from a feminist perspective by many theorists, most notably Donna Haraway and Karen Barad.

Haraway takes up Butler's notion of *materialization*. In *Bodies that Matter*, the author writes: “the notion of matter [should not be considered] as site or surface, but as a process of materialization that stabilizes over time to produce the effect of boundary, fixity, and surface we call matter” [65, p. 10]. The body is not a neutral surface on which cultural meanings are recorded. Rather, it is the result of a process of constant production of its boundaries. Think, for example, of technologies that make it possible to visualize the fetus in the womb. They produce a certain materialization of the child's gendered body, inscribing it into a binary symbolic framework even before birth [66].

In this regard, Donna Haraway speaks of *body production apparatuses* to indicate the process by which the boundaries of the object of knowledge are established in the interaction between different actors. She writes: “bodies as objects of knowledge are materialsemiotic generative nodes. Their boundaries materialize in social interaction among humans and non-humans, including the machines and other instruments that mediate exchanges at crucial interfaces and that function as delegates for other actors' functions and purposes”. [67, p. 298].

Finally, Karen Barad shows how every process of measurement in scientific practice is a device for the production of meaning. She writes: “the measurement apparatus is the condition of possibility for determinate meaning for the concept in question [e.g. gender], as well as the condition of possibility for the existence of determinately bounded and propertied” [33, p. 128].

Thus, performative theory, as used by these authors, shows how the phenomena of discrimination are complex and involve processes of symbolic meaning production. The latter emerges from the nexus of human, non-human and technological components [59].

Understanding the discrimination, the symbolic and material violence that materializes in the nodes that shape our societies means working on the creation of those nodes. It means adopting a diffractive perspective, i.e. dislocating the spokes that connect the elements of a network in relationships of domination and oppression [67, p. 302]. From these perspectives, the investigation of the union between the human and the technical emerges. Rather than being conceived of as ontologically distinct, these two dimensions are understood as part of a processual continuum. It is precisely from this continuum that the process of symbolic production can be analyzed, thus enabling an assessment of the structure of human-AI relations.

5. Examining the COMPAS Case Using Performative Theory

In this section, we briefly try to show how it is therefore possible to use the gender performativity theory to re-read one of the most exemplary cases of bias in ML that has emerged in recent years: The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software developed by Northpoint. This case has not only attracted the attention of specialists but has also gained considerable traction in the public debate. It is precisely for this reason that we have chosen to analyze it, as we believe that the perspective of gender performativity allows us to bring to light

some novel aspects, thus enriching the debate on discrimination in AI. Furthermore, it is evident that intersectionality, defined as the interaction between multiple forms of discrimination at the societal level, is a prominent factor in this case. The interconnection between gender and race dimensions is particularly noteworthy. We show that they appear to reinforce each other in a dynamic that can be described as a form of performative reinforcement.

The COMPAS software, developed in the early 2000s, was designed with the objective of calculating the probability of recidivism for a given defendant. The measure has been adopted in several states, including New York, Wisconsin, and California [68].

The software employs a variety of indicators from the subject's past, including a history of violence, substance abuse, and social environment, to categorize them according to a criminal typology [69]. This enables the prediction of the probability of future violent or law-breaking behavior. The program was trained using a dataset comprising over 30,000 samples, which were collected between 2004 and 2005 as part of a company-wide initiative involving prisons, probation, and parole facilities across the United States [70]. From this data set, the programmers identified two primary categories of criminal behavior, differentiated by gender, which are further subdivided into distinct subcategories [69]. For this reason, we have the following typologies, 8 for each gender:

	Male Typology	**Female Typology**
1.	Chronic drug abusers – most non-violent	Drug problems and anti-social sub-cultural influences – some with relationship conflicts
2.	Low risk situational – fighting/domestic violence caution	Family disorganization and inadequate parenting – residential instability and minor non-violent offences
3.	Chronic alcohol problems – DUI, domestic violence	Chronic substance abusers – women with higher social resources than other groups
4.	Socially marginalized – poor, uneducated, stressed, habitual offenders	Marginalized poor and isolated older women – economic survival crimes
5.	Criminally versatile – young marginalized persons often gang affiliated	Young antisocial poorly educated women with some violent offences and early delinquency onset
6.	Socially isolated long term substance abuse – multiple minor and mostly non-violent offenses	Chronic long term criminal history A – multiple co-occurring social and psychological risk factors
7.	Serious versatile high risk individuals	Chronic long term criminal history B – multiple co-occurring problems and high risk
8.	Low risk situational accidental category	Late starters with multiple strengths and fewer risk factors – minor non-violent offence history

In 2016, the independent editorial office ProPublica initiated an investigation to ascertain the degree of reliability of COMPAS' predictions. The findings revealed that the overall accuracy of the results was approximately 63.3%. Of greater significance was the observation that individuals identified as Black were 77% more likely to be classified as high-risk and to perpetrate a future criminal act [71]. The data, which is truly staggering in its implications, revealed that the software was inherently affected by the presence of biases that generated processes of racial discrimination. It was observed that the program is unable to meet the demand for fairness, as it is incapable of ensuring the generation of a fair output regarding all social groups [72], [73], [74]. The statistical parity component is not met due to the inadequacy of the data utilized for training the program, which contributes to the reinforcement and propagation of racial stereotypes that render Black subjectivities the most susceptible to criminality.

From a performative perspective, the problem of fairness can be rephrased as follows: *If, as it has been demonstrated previously through the application of causal and counterfactual approaches, the aspect of fairness is guaranteed when certain sensitive features do not compromise the result produced by the algorithm, then why are certain features, such as gender and race, decisive in this specific context?*

The COMPAS case illustrates the necessity of viewing gender and race not as individual attributes, but as inscribed in a broader institutional and judicial context. In this sense, COMPAS is configured precisely as an apparatus of bodily production, whereby bodies are materialized in accordance with specific codes. In this case, the production of the body follows the reiteration and citation of some isolated normative patterns in the sixteen typologies that match male and female subjects. This indicates that the algorithm performs a process of constructing criminal subjectivity, which is intimately connected to the normative criteria. The sixteen proposed typifications thus become the normative lenses through which the software produces its judgment. The concepts of gender and race do not exist in and of themselves; rather, they are constituted through a process of materialization that produces and naturalizes certain subjectivities on the basis of specific extrinsic characteristics, they are intersected. In this sense, the process of racialization is perpetrated and repeated within the framework of the long institutional and legal history of violence against subjectivities of color [75], [76]. The COMPAS system establishes a norm based on the historical datasets through which it was trained, thereby reproducing the observed effects. In this sense, it can be seen to contribute to an epistemic injustice that criminalizes minority subjectivities, such as those of women of color and immigrants, by automating and thereby making this process more efficient. It reinforces the existing discriminatory social norms embedded in the American legal system, thereby contributing to the creation of an inequitable and undemocratic network. It is therefore necessary to connect the social ontology in which these actors—software, the U.S. institutional legal apparatus, and so on—can redefine the instance of fairness [77], [78]. In this case, gender and race cannot simply be regarded as sensitive categories; rather, ways must be found through which the computational power of ML can be employed to modify the constituent elements of the network to break down processes of sexual and racial discrimination. This approach thus enables not only the diagnosis of the way disparate social actors are connected, but also the comprehension of the processes of signification attributed to the bodies within this mechanism. As stated in the EDRi report "Verifying that a system is fair with the current focus on models' outputs is, then, not enough, as we also need to analyze the negative impact the new system might have on the entire, original environment" [25, p. 63]. For this reason, an intersectional approach to algorithmic fairness can be achieved if we frame it in a previous consideration on how the network is established and which processes of performative signification it implies.

6. Conclusion

As can be seen from the discussion above, it is difficult to confine the issue of gender or even racial discrimination to the technical sphere. In fact, AI in general, and ML in particular, fits perfectly into the mechanism of meaning production described above [79]. In fact, Hoffmann writes: "algorithms do not merely shape distributive outcomes, but they are also intimately bound up in the production of particular kinds of meaning, reinforcing certain discursive frames over others" [80, p. 908]. It is desirable to make software as free from bias as possible, but this alone is not enough to contrast discrimination.

Discriminatory phenomena arise from the intertwining of human and technological components. This is why the causal and counterfactual reasoning used to guarantee the fairness of a program is not sufficient [81]. Gender is not an attribute and reality in itself, but the spectrum against which we measure the way in which relationships between humans, non-humans, technologies and the environment are structured according to power relations.

The discussion on fairness should therefore not only consider gender as an attribute or a sensitive characteristic to be managed but must also consider the social ontology from which it emerges, becomes problematic and produces an asymmetrical relationship between people. Kohler-Hausmann

invites us to make the same argument in relation to processes of racialization, writing: "We often lose sight of the practices and meanings that constitute the very categories of race because one of the properties of this social category is to appear as a natural fact about bodies instead of the effect of persistent social stratification and meaning-making" [82, p. 1225].

ML has a powerful capacity for agency and can therefore prove to be an excellent tool for modifying the many social conditions that make the concept of gender relevant in a given context. However, this is an operation that must be carried out from time to time for each type of program that will be developed [59].

It is therefore still necessary to understand how the gender dimension is constructed and used from time to time in ML, as Rode invites us to do with the concept of *gender position* [83]. She shows how every technological innovation, such as the widespread diffusion of household appliances, is always accompanied by a redefinition of social roles.

In this sense, this study tries to underscore the critical need to address gender bias in AI systems through a multifaceted approach that integrates technical rigor with a deep understanding of social dynamics. For this reason, based on the findings, three recommendations can be made in order to build more equitable AI systems:

1. **Promote Interdisciplinary Collaboration:** First, it is important to highlight the importance of creating an interdisciplinary program able to bridge computer science with the humanities. AI developers should work closely with social scientists, ethicists, and feminist scholars to ensure that ethical considerations, especially those relating to gender and intersectionality, are embedded in AI development processes.
2. **Redefine AI Ethics to Incorporate Social Ontology:** To this end, AI ethics should be expanded beyond technical fairness metrics to incorporate the discussion on how social ontology is constructed. This involves, as the first step, recognizing that gender is not a fixed attribute but a social construct that influences how AI systems are developed and deployed.
3. **Continuous Monitoring and Evaluation of AI Systems:** It is recommended to implement continuous monitoring and post-implementation evaluation mechanisms for AI systems to identify and correct any discriminatory effects that may emerge over time. This could be facilitated by establishing independent ethics committees that regularly assess the operation of AI systems going beyond the actual procedures of auditing and debiasing.

We aim to contribute to the field by bridging the gap between technical AI research and critical Gender Studies, offering a comprehensive framework for understanding and addressing intersectional biases. This contribution lays the groundwork for future research that further explores the intersections of AI, gender and social justice, advancing the development of more equitable AI technologies.

Acknowledgements

This work was partially supported by the project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU).

References

[1] T. M. Mitchell, *Machine Learning*. in McGraw-Hill series in computer science. New York: McGraw-Hill, 1997.

[2] L. Floridi, *The ethics of artificial intelligence: principles, challenges, and opportunities*. New York: Oxford University Press, 2023.

[3] N. K. Hayles, "Inside the Mind of an AI: Materiality and the Crisis of Representation," *nlh*, vol. 54, no. 1, pp. 635–666, Sep. 2022, doi: 10.1353/nlh.2022.a898324.

[4] N. K. Hayles, *Unthought: the power of the cognitive nonconscious*. Chicago ; London: The University of Chicago Press, 2017.
[5] H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, -- NY USA: ACM, Oct. 2021, pp. 1–9. doi: 10.1145/3465416.3483305.
[6] European Union Agency for Fundamental Rights, "#BigData: Discrimination in data-supported decision making," 2018.
[7] European Union Agency for Fundamental Rights, "Bias in algorithms - Artificial intelligence and discrimination," 2022.
[8] European Union Agency for Fundamental Rights, "Getting the future right – Artificial intelligence and fundamental rights," 2020. [Online]. Available: https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights
[9] S. Hekman, "Beyond identity: Feminism, identity and identity politics," *Feminist Theory*, vol. 1, no. 3, pp. 289–308, Dec. 2000, doi: 10.1177/14647000022229245.
[10] P. Bourdieu, *Masculine domination*, 1. publ. Cambridge: Polity Press, 2001.
[11] P. Gilroy, *The black Atlantic: modernity and double consciousness*, 8. print. Cambridge, Mass: Harvard Univ. Press, 2003.
[12] L. Irigaray, *Speculum of the other woman*. Ithaca, N.Y: Cornell University Press, 1985.
[13] R. Braidotti, *The posthuman*. Cambridge, UK: Polity Press, 2013.
[14] A. Cavarero, *Inclinations: a critique of rectitude*. in Square one. Stanford (Calif.): Stanford university press, 2016.
[15] S. de Beauvoir, C. Capisto-Borde, and S. Malovany-Chevallier, *The second sex*, First Vintage Books ed. New York: Vintage Books, 2011.
[16] J. Butler, *Gender trouble: feminism and the subversion of identity*. in Routledge classics. New York: Routledge, 2006.
[17] C. Bottici, *Anarchafeminism*. London ; New York: Bloomsbury Academic, 2022.
[18] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: limitations and opportunities*. Cambridge, Massachusetts: The MIT Press, 2023.
[19] D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Computers in Human Behavior*, vol. 98, pp. 277–284, Sep. 2019, doi: 10.1016/j.chb.2019.04.019.
[20] J. C. Bjerring and J. Busch, "Artificial intelligence and identity: the rise of the statistical individual," *AI & Soc*, Mar. 2024, doi: 10.1007/s00146-024-01877-4.
[21] S. T. J. Hudson, A. Myer, and E. C. Berney, "Stereotyping, prejudice, and discrimination at the intersection of race and gender: An intersectional theory primer," *Social & Personality Psych*, vol. 18, no. 2, Feb. 2024, doi: 10.1111/spc3.12939.
[22] P. Hill Collins, *Intersectionality*, Second edition. in Key concepts. Cambridge Medford, Mass: Polity press, 2020.
[23] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," in *Feminist legal theories*, Routledge, 2013, pp. 23–51.
[24] R. Islam, K. N. Keya, S. Pan, A. D. Sarwate, and J. R. Foulds, "Differential Fairness: An Intersectional Framework for Fair AI," *Entropy*, vol. 25, no. 4, p. 660, Apr. 2023, doi: 10.3390/e25040660.
[25] A. Balayn and S. Gürses, "Beyond Debiasing: Regulating AI and its inequalities," European Digital Rights (EDRi), 2021. [Online]. Available: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf
[26] L. Floridi, "Technology's In-Betweeness," *Philos. Technol.*, vol. 26, no. 2, pp. 111–115, Jun. 2013, doi: 10.1007/s13347-013-0106-y.
[27] M. Dolata, S. Feuerriegel, and G. Schwabe, "A sociotechnical view of algorithmic fairness," *Information Systems Journal*, vol. 32, no. 4, pp. 754–818, Jul. 2022, doi: 10.1111/isj.12370.
[28] L. A. Suchman, *Plans and situated actions: the problem of human-machine communication*, Nachdr. in Learning in doing: social, cognitive, and computational perspectives. Cambridge: Cambridge Univ. Press, 1999.
[29] B. Latour, *Reassembling the social: an introduction to Actor-Network-Theory*, 1. publ. in pbk. in Clarendon lectures in management studies. Oxford: Oxford Univ. Press, 2007.
[30] B. Latour, "On technical mediation," *Common knowledge*, vol. 3, no. 2, pp. 29–64, 1994.

[31] K. Barad, “Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter,” *Signs: Journal of Women in Culture and Society*, vol. 28, no. 3, pp. 801–831, Mar. 2003, doi: 10.1086/345321.
[32] W. J. Orlikowski, “Sociomaterial Practices: Exploring Technology at Work,” *Organization Studies*, vol. 28, no. 9, pp. 1435–1448, Sep. 2007, doi: 10.1177/0170840607081138.
[33] K. M. Barad, *Meeting the universe halfway: quantum physics and the entanglement of matter and meaning*. Durham: Duke University Press, 2007.
[34] M. Coeckelbergh, *Why AI undermines democracy and what to do about it*. Medford: Polity Press, 2024.
[35] B. Stiegler and D. Ross, *What makes life worth living: on pharmacology*, English edition. Cambridge, UK: Polity, 2013.
[36] Plato and J. Derrida, *Phèdre*, Nouvelle édition corrigée et mise à jour. Paris: GF Flammarion, 2004.
[37] E. Ferrara, “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies,” *Sci*, vol. 6, no. 1, p. 3, Dec. 2023, doi: 10.3390/sci6010003.
[38] C. S. Calude and G. Longo, “The Deluge of Spurious Correlations in Big Data,” *Found Sci*, vol. 22, no. 3, pp. 595–612, Sep. 2017, doi: 10.1007/s10699-016-9489-4.
[39] X. Chang, “Gender Bias in Hiring: An Analysis of the Impact of Amazon’s Recruiting Algorithm,” *AEMPS*, vol. 23, no. 1, pp. 134–140, Sep. 2023, doi: 10.54254/2754-1169/23/20230367.
[40] F. Pasquale, *The black box society: the secret algorithms that control money and information*, First Harvard University Press paperback edition. Cambridge, Massachusetts London, England: Harvard University Press, 2016.
[41] M. Bunge, “A General Black Box Theory,” vol. 30, no. 4, pp. 346–358, 1963.
[42] W. H. K. CHUN, *DISCRIMINATING DATA: correlation, neighborhoods, and the new politics of recognition*. S.l.: MIT PRESS, 2024.
[43] P. Boddington, “The Rise of AI Ethics,” in *AI Ethics*, in Artificial Intelligence: Foundations, Theory, and Algorithms. , Singapore: Springer Nature Singapore, 2023, pp. 35–89. doi: 10.1007/978-981-19-9382-4_2.
[44] R. Dwivedi *et al.*, “Explainable AI (XAI): Core Ideas, Techniques, and Solutions,” *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.
[45] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, “Algorithmic fairness datasets: the story so far,” *Data Min Knowl Disc*, vol. 36, no. 6, pp. 2074–2152, Nov. 2022, doi: 10.1007/s10618-022-00854-z.
[46] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.
[47] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges,” in *Natural Language Processing and Chinese Computing*, vol. 11839, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., in Lecture Notes in Computer Science, vol. 11839. , Cham: Springer International Publishing, 2019, pp. 563–574. doi: 10.1007/978-3-030-32236-6_51.
[48] A. Ryan, “Fairness and Philosophy,” *Social Research*, vol. 73, no. 2, pp. 597–606, 2006.
[49] J. Rawls, *Justice as fairness: a restatement*, 3. printing. Cambridge, Mass.: Belknap Press of Harvard University Press, 2003.
[50] J. Broome, “V—Fairness,” *Proceedings of the Aristotelian Society*, vol. 91, no. 1, pp. 87–102, Jun. 1991, doi: 10.1093/aristotelian/91.1.87.
[51] R. Van Nood and C. Yeomans, “Fairness as Equal Concession: Critical Remarks on Fair AI,” *Sci Eng Ethics*, vol. 27, no. 6, p. 73, Dec. 2021, doi: 10.1007/s11948-021-00348-z.
[52] J. M. Bishop, “Artificial Intelligence is stupid and causal reasoning won’t fix it,” 2020, *arXiv*. doi: 10.48550/ARXIV.2008.07371.
[53] P. J. Bickel, E. A. Hammel, and J. W. O’Connell, “Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.,” *Science*, vol. 187, no. 4175, pp. 398–404, Feb. 1975, doi: 10.1126/science.187.4175.398.
[54] K. H. Chu, N. J. Brown, A. Pelecanos, and A. F. Brown, “Simpson’s paradox: A statistician’s case study,” *Emerg Medicine Australasia*, vol. 30, no. 3, pp. 431–433, Jun. 2018, doi: 10.1111/1742-6723.12943.

[55] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*, First edition. New York: Basic Books, 2018.
[56] J. Pearl, "CAUSALITY: MODELS, REASONING, AND INFERENCE, by Judea Pearl, Cambridge University Press, 2000," *Econ. Theory*, vol. 19, no. 04, Aug. 2003, doi: 10.1017/S0266466603004109.
[57] M. Ziosi, D. Watson, and L. Floridi, "A Genealogical Approach to Algorithmic Bias," *SSRN Journal*, 2024, doi: 10.2139/ssrn.4734082.
[58] L. Hu and I. Kohler-Hausmann, "What's sex got to do with machine learning?," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona Spain: ACM, Jan. 2020, pp. 513–513. doi: 10.1145/3351095.3375674.
[59] E. Drage and F. Frabetti, "AI that Matters: A Feminist Approach to the Study of Intelligent Machines," in *Feminist AI*, 1st ed., J. Browne, S. Cave, E. Drage, and K. McInerney, Eds., Oxford University PressOxford, 2023, pp. 274–289. doi: 10.1093/oso/9780192889898.003.0016.
[60] G. Rubin, "The Traffic in Women: Notes on the 'Political Economy' of Sex," in *Deviations*, Duke University Press, 2012, pp. 33–65. doi: 10.1215/9780822394068-002.
[61] D. Bachmann-Medick, "Chapter II: The Performative Turn," in *Cultural Turns*, De Gruyter, 2016, pp. 73–102. doi: 10.1515/9783110402988-004.
[62] C. Licoppe, "THE 'PERFORMATIVE TURN' IN SCIENCE AND TECHNOLOGY STUDIES: Towards a linguistic anthropology of 'technology in action,'" *Journal of Cultural Economy*, vol. 3, no. 2, pp. 181–188, Jul. 2010, doi: 10.1080/17530350.2010.494122.
[63] P. S. Jasanoff, D. G. E. E. Markle, J. C. C. Peterson, and D. T. J. Pinch, *Handbook of Science and Technology Studies*. Thousand Oaks: SAGE Publications, 2001.
[64] B. Latour, *We have never been modern*. Cambridge, Massachusetts: Harvard University Press, 1993.
[65] J. Butler, *Bodies that matter: on the discursive limits of "sex."* in Routledge classics. Abingdon, Oxon ; New York, NY: Routledge, 2011.
[66] J. Butler, *Undoing gender*. New York ; London: Routledge, 2004.
[67] D. J. Haraway, *Manifestly Haraway*. Minneapolis: University of Minnesota Press, 2016.
[68] K. Kirkpatrick, "It's not the algorithm, it's the data," *Commun. ACM*, vol. 60, no. 2, pp. 21–23, Jan. 2017, doi: 10.1145/3022181.
[69] Northpoint, "Practitioner's Guide to COMPAS Core." 2015. [Online]. Available: https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf
[70] M. A. Vaccaro, "Algorithms in human decision-making: A case study with the COMPAS risk assessment software," 2019.
[71] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "How We Analyzed the COMPAS Recidivism Algorithm." 2016.
[72] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Sci. Adv.*, vol. 4, no. 1, p. eaao5580, Jan. 2018, doi: 10.1126/sciadv.aao5580.
[73] F. Gursoy and I. A. Kakadiaris, "Equal Confusion Fairness: Measuring Group-Based Disparities in Automated Decision Systems," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Orlando, FL, USA: IEEE, Nov. 2022, pp. 137–146. doi: 10.1109/ICDMW58026.2022.00027.
[74] F. Lagioia, R. Rovatti, and G. Sartor, "Algorithmic fairness through group parities? The case of COMPAS-SAPMOC," *AI & SOCIETY*, vol. 38, no. 2, pp. 459–478, 2023.
[75] S. Browne, *Dark Matters: On the Surveillance of Blackness*. Duke University Press, 2015, p. dup;9780822375302/1. doi: 10.1215/9780822375302.
[76] A. Shapiro, "Reform predictive policing," *Nature*, vol. 541, no. 7638, pp. 458–460, Jan. 2017, doi: 10.1038/541458a.
[77] T. R. Clemons, "Blind injustice: The Supreme Court, implicit racial bias, and the racial disparity in the criminal justice system," *Am. Crim. L. Rev.*, vol. 51, p. 689, 2014.
[78] I. M. Young, *Justice and the politics of difference*, Nachdr. Princeton, NJ: Princeton Univ. Press, 1990.
[79] E. Drage and F. Frabetti, "The Performativity of AI-powered Event Detection: How AI Creates a Racialized Protest and Why Looking for Bias Is Not a Solution," *Science, Technology, & Human Values*, p. 016224392311646, Mar. 2023, doi: 10.1177/01622439231164660.

[80] A. L. Hoffmann, "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, vol. 22, no. 7, pp. 900–915, Jun. 2019, doi: 10.1080/1369118X.2019.1573912.
[81] S. Ruggieri, J. M. Alvarez, A. Pugnana, L. State, and F. Turini, "Can We Trust Fair-AI?," *AAAI*, vol. 37, no. 13, pp. 15421–15430, Jun. 2023, doi: 10.1609/aaai.v37i13.26798.
[82] I. Kohler-Hausmann, "The Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination," *SSRN Journal*, 2017, doi: 10.2139/ssrn.3050650.
[83] J. A. Rode, "A theoretical agenda for feminist HCI," *Interacting with Computers*, vol. 23, no. 5, pp. 393–400, Sep. 2011, doi: 10.1016/j.intcom.2011.04.005.