OPEN ACCESS

*Sexuality and Gender Studies Journal*
ISSN 2783-7130

# Exploring the Question of Bias in AI through a Gender Performative Approach

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

[1] DIRIUM Department, University of Bari Aldo Moro, Italy
[2] Department of Informatics, University of Bari Aldo Moro, Italy

**ARTICLE INFO**

***Keywords:***
*Algorithmic Discrimination,*
*Gender bias,*
*Performative theory,*
*AI ethics,*
*Fairness in AI*

**ABSTRACT**

The objective of this paper is to examine how artificial intelligence systems (AI) can reproduce phenomena of social discrimination and to develop an ethical strategy for preventing such occurrences. A substantial body of scholarship has demonstrated how AI has the potential to erode the rights of women and LGBT+ individuals, as it is capable of amplifying forms of discrimination that are already pervasive in society. This paper examines the principal approaches that have been put forth to contrast the emergence of biases in AI systems, namely causal, counterfactual reasoning, and constructivist methodology. This analysis demonstrates the necessity of considering the sociopolitical context in which AI systems are developed when evaluating their ethical implications. To investigate this conjunction, we apply the theory of gender performativity as theorized by Judith Butler and Karen Barad. 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. In conclusion, we demonstrate how reframing of gender performativity theory, when applied to AI ethics, permits us to consider the social context within which these technologies will operate. This approach enables an expansion of the interpretation of the concept of fairness, thereby reflecting the complex dynamics of gender production. In the context of AI ethics, the concept of "fairness" pertains to the capacity of an algorithm to generate results dealing with sensitive categories, such as gender, ethnicity, religion, sexual orientation, and disability, in a manner that does not engender forms of discrimination and prejudice. 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 development, testing, and evaluation phases of AI systems.

[*] Corresponding author's E-mail address: gabriele.nino@uniba.it

**Cite this article as:**
Nino, G., & Lisi, F. A. (2024). A Performative Approach for Rethinking the Question of Gender Bias in AI. *Sexuality and Gender Studies Journal*, *2*(2): 14-31. https://doi.org/10.33422/sgsj.v2i2.735

## 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. In general, it refers to the use of algorithms trained on large amounts of data to perform specific tasks, especially classification and prediction (Mitchell, 1997).

The interest in these computational technologies stems mainly from the fact that they are the first form of technology ever developed that has a large reserve of agency and autonomy (Floridi, 2023). Consider, for example, the famous case of Chat GPT, developed by Open-AI, a generative system capable of performing certain tasks by understanding natural language (Hayles, 2022). 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 (Hayles, 2017). As a result, ML is opening a wide range of applications, from medicine to architecture, from engineering to finance, and so on.

Without indulging in a state of absolute celebration and idolatry towards the field of AI, many scholars have noted how the transformation of the social fabric is in parallel producing a consequential transformation of ancient relations of force and power, which in turn produce phenomena of racial and sexual discrimination (Benjamin, 2020; Eubanks, 2019; O'Neil, 2016). AI research is not only being developed from a technical perspective. AI technologies also have an inherent social and political value (Crawford, 2021). The objective is to comprehend the way social phenomena of discrimination, e.g. gender discrimination or racism, are embedded in ML and to identify an ethical concept that can be developed to circumvent this issue. In particular, we examine the notion of fairness from a feminist perspective to enhance the discourse on AI ethics, integrating insights on gender and acknowledging its complexities. To this end, we try to connect AI research with the body of feminist thought. We propose extending the theory of gender performativity, developed by Judith Butler and Karen Barad, to elucidate the propagation of gender discrimination in ML.

### 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) (Dwivedi et al., 2023; Miller, 2019; Xu et al., 2019). 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, and applying it to the case of AI. This is done to integrate the social dimension within the sphere of fairness. Based on these theoretical premises, we examine in Section 5 the COMPAS case, a software program utilized for calculating the probability of recidivism of a defendant, as an illustrative example of this method. In addition, we enucleate 5 practical principles and

recommendations that should be considered and further explored for their implications in the field of AI ethics.

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

## 2. Theoretical Framework: ML's Ethical Problems

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 can 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 performance?

First, the ML process can be represented as a loop (Barocas et al., 2023). 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. ML, by its very nature, operates at two levels: the purely technical and the social (Floridi, 2013). Therein lies its agency (Dolata et al., 2022).

However, the social sphere is made up of a series of unequal and discriminatory relationships between individuals. 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 (Coeckelbergh, 2024). For this reason, 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 (2013). In ancient Greek, *pharmakon* means both poison and remedy (Plato & Derrida, 2004). More generally, technology can be a means of promoting social justice or destroying human bonds, as in the case of the Holocaust or the atomic bomb (Anders, 2002; Arendt et al., 2018; Jonas, 1984). So how does AI fit into this dynamic?

Each of the stages in the ML lifecycle can be affected by measurement or modeling errors that lead to the production of inaccurate results. These are commonly referred to as biases. Ferrara (2023) 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. It is well-known how ML programs have extracted spurious forms of correlations between certain features from big data sets (Calude & Longo, 2017). 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 (Chang, 2023).

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 (Pasquale, 2016).

In the ’60s the Argentine-Canadian philosopher Mario Bunge 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» (1963, p. 346)

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 (Boddington, 2023). 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. Methodology

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 fairness within the field of AI ethics as a means of addressing the issue of discrimination. We begin by tracing the history of this concept, which did not originate in the field of computer science. We analyze from the qualitative viewpoint the most significant techniques through which the concept of fairness has been formalized.

To identify the relevant literature, we have conducted a comprehensive search of peer-reviewed articles and books. The databases used include Google Scholar, Scopus, and IEEE Xplore, focusing on sources published within the last ten years to ensure relevance to contemporary debates. The search was conducted using three different keywords namely ‘gender bias in AI,’ ‘algorithmic fairness,’ and ‘performative theory in AI,’ employing also Boolean operators to refine the search process. The results of this process allowed for the identification of the most influential contributions on algorithmic fairness, which were selected based on specific exclusion criteria. To ensure the quality and relevance of the sources, we excluded those that lacked methodological rigor, were outdated, or did not specifically address gender or racial discriminatory dynamics in AI systems. In contrast, we included all contributions explicitly addressing gender bias in ML from a standpoint encompassing both computer science and feminist theory. This approach allowed us to examine how bias is addressed in different academic fields. Our analysis then evaluated the qualitative aspects, specifically identifying

and synthesizing the most prevalent methods used to contrast discrimination in the field of computer science.

In this regard, we have identified causal and counterfactual reasoning 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 comparative approach that can enable us to connect these two areas of research.

### 3.1. Historical Aspects: Fairness in ML

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 (Shin & Park, 2019). These principles have become hegemonic in the field of so-called XAI (Dwivedi et al., 2023; Fabris et al., 2022; Miller, 2019; Xu et al., 2019). 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 (Ryan, 2006). 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* (2003). Since the 1990s, the concept of fairness has also been used in sociology and economics (Broome, 1991). 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 (Van Nood & Yeomans, 2021).

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 the ethics of AI.

### 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” (Bishop, 2020). 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 admission rates at the University of Berkeley in 1973 (Bickel et al., 1975).

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 (Chu et al., 2018).

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» (2018, 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 (Pearl & Mackenzie, 2018, 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 (Miller, 2019).

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?” (Pearl, 2003).

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 leveled at this theoretical framework.

First, Ziosi et al. showed how, in XAI, the notion of fairness is measured only *a posteriori*, i.e. based on 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 (2024).

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 (2020). 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" (2024).

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 (Drage & Frabetti, 2023a).

## 4. Research Findings and Discussion

### 4.1. Feminism's Insight into the Gender Dimension

Since the second wave of feminism, it has been shown that gender difference cannot be reduced to a purely biological factor. Rather, it has been examined in terms of how human relations have taken on the differences between bodies by ontologizing them on a scale of values. The patriarchal form adopted by Western societies has taken as its normative ideal the neutral subject as theorized by René Descartes and the European Enlightenment. The neutral subject seems to have no gendered body, which only theoretically guarantees its universal status. All forms of subjectivity can thus be reflected in it (Irigaray, 1985).

But what works on a theoretical level then suffers from the devastating effects of the reality principle. Indeed, it is well known how early theorists in the history of feminist thought, such as Mary Wollstonecraft or Olympe de Gouges, denounced the non-neutrality of this model, of a universality that instead surreptitiously prescribes a precise social ideal as normative (Hirschmann & Regier, 2019; Lloyd, 1979; Siess, 2005). Under the guise of the neutral and universal Enlightenment subject lies the figure of the Western, heterosexual white man (Braidotti, 2013). All other subjectivities that deviate or tend to deviate are disqualified and subjected to symbolic violence. This is the root of patriarchal and colonial violence. To deviate from the universal model is to be worth less than the hegemonic subject (Cavarero, 2016; De Lauretis, 1990; Mbembe, 2020; Sedgwick, 1993).

The power of this symbolic process lies in the assumption of extrinsic and phenotypic differences between subjectivities, such as female genitalia, skin color, and sexual orientation, as an index of social inferiority. The subjectivities bearing these stigmas are socially signified by a negative symbolic capital. For this reason, their social position is radically constructed. But by anchoring itself in bodily difference, the process of social construction conceals its artificiality by masquerading it as natural difference. The French sociologist Pierre Boudieu has described this process admirably. He writes: «Whatever their position in the social space, women have in common the fact that they are *separated from men by a negative symbolic coefficient* which, like skin color for blacks, or any other sign of membership of a stigmatized group, negatively affects everything that they are and do» (Bourdieu, 2001, p. 93). Bourdieu therefore invites us to look at social cosmology, i.e. the set of social arrangements *introjected* by social actors, to understand how forms of discrimination are constructed.

Italian feminist and lesbian philosopher Teresa de Lauretis speaks of, borrowing the term from Michel Foucault's philosophy, "technologies of gender" (De Lauretis, 1987). With this concept, the author wants to show how gender is the result of the action of socio-political techniques and biomedical apparatuses that shape subjectivity. For this reason, she writes: "a subject constituted in gender […] though not by sexual difference alone, but rather across languages and cultural representations" (ivi, p.2).

Finally, Judith Butler develops her performative perspective by showing that gender is the result of a process of citation of certain social norms (Butler, 2006).

Michel Foucault, Teresa de Lauretis, and Judith Butler have all shown, in different ways, how sex or gender is socially constructed using certain *social technologies* (Behrent, 2013; Dişci, 2024). What is surprising, however, is that these authors rarely analyze how the technologies themselves contribute to the construction of gender. For example, Foucault analyses the process by which homosexuality was medicalized in the 19th century (Foucault, 2008). De Lauretis shows how cinematic imagery constructs the figure of the lesbian (De Lauretis, 1984, 1987, 1994), and Butler examines how linguistic techniques give stability to gender identities (Butler, 2004, 2006, 2011).

But as we have shown above, new digital technologies, including AI, have profoundly altered the social fabric. Thus, it is necessary to rethink the dynamics of gender construction across the spectrum of these technologies in order to provide a solid foundation of social ontology on which to base AI ethics.

This does not mean abandoning previous theories. On the contrary, we believe that the theory of performativity needs to be reformulated and modified to become a guide for the present. In what follows, we attempt to carry out this work of revision.

### 4.2. The Origin of Gender Performative Theory

Since the 1990s, the American philosopher Judith Butler has proposed to re-read the question of sexual difference through a performative lens.

The theory of performativity has its origins in the linguistic research of John L. Austin (Austin & Urmson, 2009). He showed how linguistic acts are not merely descriptive utterances but produce effects in the real world. For example, the utterance 'I pronounce you man and wife' functions like a magic formula: it has the power to transport the referent subjects of the utterance to another normative level, changing their social status (Butler & Shusterman, 1999). But where does the normative power of a performative utterance come from?

Butler is inspired by a Derridean reinterpretation of the Kafkaesque novella 'Before the Law'. This parable tells the story of a subject who is interpellated and determined by an anonymous law without origin (Cornell et al., 2016; Derrida, 2018). For Butler, this is the performative dynamic that constructs the sexual subject (Butler, 1997). There is a symbolic law that determines subjects by providing a stable and coherent identity in return. But what does this process involve?

In *Gender Trouble* (Butler, 2006), 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 (Rubin, 2012). 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” (Butler, 2006, p. 11).

This ordinary conception is 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. For Butler, following Foucault, on a diachronic level, it is not

possible to identify a temporal origin from which symbolic law is exercised, much less, on a synchronic level, to imagine a region outside the normative power of law. 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. Gender, 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" (ivi, p. 33).

Thus, gender is not an individual attribute of human beings, but rather 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.

### 4.3. Performativity in Science and Technology Studies (STS)

Beginning with the pioneering work of Judith Butler, the theme of the performative construction of the categorical framework through which we usually think about our social experience has been taken seriously by various authors (Bachmann-Medick, 2016; Licoppe, 2010). In different fields, performance theory has been used to deconstruct the forms of power that are unduly repeated in them. It has also found fertile ground to take root in the field of Science and Technology Studies (STS). 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 (Jasanoff et al., 2001).

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 (Latour, 1993). From a Kantian perspective, for example, the phenomenon is that which manifests itself to the transcendental subject through its categorical apparatus. The 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 (Latour, 2007). 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" (2011, 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 (Butler, 2004).

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". (2016, 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" (2007, 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 (Drage & Frabetti, 2023a).

Understanding the discrimination, and 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 (Haraway, 2016, p. 302).

## 5. Applying Performative Theory to AI Ethics

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. In light of the aforementioned example, we enumerate five generic guidelines that should be kept in mind when discussing the concept of fairness in AI ethics.

### 5.1. Examining the COMPAS Case Using Performative Theory

The COMPAS software, developed in the early 2000s, was designed to calculate the probability of recidivism for a given defendant. 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 (Northpoint, 2015). 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 (Vaccaro, 2019). From this data set, the programmers identified two

primary categories of criminal behavior, differentiated by gender, which are further subdivided into distinct subcategories. The male typology is comprised of eight categories, including "chronic drug abuser," "socially marginalized," "criminally versatile," and others. The female typology is similarly divided into eight categories, with examples such as "chronic long-term criminal history" and "young antisocial, uneducated women"(Northpoint, 2015). The measure has been adopted in several states, including New York, Wisconsin, and California (Kirkpatrick, 2017).

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 (Angwin et al., 2016). 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 (Dressel & Farid, 2018; Gursoy & Kakadiaris, 2022; Lagioia et al., 2023). 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 based on specific extrinsic characteristics. 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 (S. Browne, 2015; Shapiro, 2017). 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 (Clemons, 2014; Young, 1990). 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.

### 5.2. Practical Advises and Recommendations

While the research presented in this study requires further investigation and development, it is possible to make several recommendations based on the findings. The findings of this study demonstrate the necessity of addressing gender bias in AI systems through a dual approach: theoretical analysis and practical steps that can be implemented by researchers, policymakers, and AI developers. The complexity of the issues uncovered suggests that a multifaceted approach is necessary to create more equitable AI systems. The following section presents a series of concrete recommendations aimed at mitigating bias and promoting fairness in AI development and deployment. These recommendations are intended to guide both future research and practical applications, ensuring that AI systems contribute to a more just and inclusive society:

1. **Promote Interdisciplinary Collaboration:** 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. Academic institutions should create more interdisciplinary research programs that bridge computer science with social justice frameworks.
2. **Redefine AI Ethics to Incorporate Social Ontology:** Theoretical frameworks for AI ethics should expand beyond technical fairness metrics to incorporate the social ontology of gender. This involves 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.
4. **Involvement of Different Stakeholders in AI Design**: It is recommended that stakeholders representing groups that may be subject to discrimination, such as women, LGBTQ+ individuals, and ethnic minorities, be included in the AI design process. This can be achieved through participatory workshops or focus groups that allow these groups to contribute directly to the development of fairer and more inclusive AI systems.
5. **Training on AI Ethics:** It is recommended that technology companies and public institutions incorporate training programs on AI ethics into their curricula for developers and researchers. These programs should emphasize the social implications of algorithmic bias and the analysis of the intersection between digital technologies and gender and racial issues. Such training would enhance awareness of the potential discriminatory effects of algorithms.

For these reasons, to effectively address these biases, it is imperative to transcend the limitations of conventional technical solutions and embrace a comprehensive approach that encompasses diverse perspectives and interdisciplinary collaboration. The proposed recommendations highlight practical measures that researchers, policymakers, and AI developers can implement to foster the development of more equitable and transparent AI systems. This entails the incorporation of stakeholders from marginalized groups in the design phase, and the establishment of continuous monitoring mechanisms to guarantee the long-term fairness and impartiality of AI systems.

Furthermore, these recommendations underscore the necessity for a comprehensive reassessment of AI ethics. It is imperative that current frameworks be expanded to address the social ontology of gender and other intersecting identities. This will ensure that AI technologies

not only avoid harm but also actively contribute to social justice. By integrating feminist perspectives and engaging in cross-disciplinary collaboration, the development of AI can be reoriented toward promoting equality rather than perpetuating existing inequalities.

## 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 (Drage & Frabetti, 2023b). 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" (2019, 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 (Ruggieri et al., 2023). Gender is not an attribute and a 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 about 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" (2017, 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 (Drage & Frabetti, 2023a). Nevertheless, the implementation of the five guidelines in the development and programming of ML systems would facilitate a more accurate perspective on gender issues, thereby enhancing comprehension of the processes through which discriminatory phenomena emerge and are perpetuated. However, further investigation is required to ascertain the potential consequences and effects that these recommendations may have on the field of software programming.

In this sense is still crucial 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* (2011). She shows how every technological innovation, such as the widespread diffusion of household appliances, is always accompanied by a redefinition of social roles. However, there is no inherent correspondence between the development of appliances and the progressive process of emancipation. The notion of gender position invites us to look beyond the ideology of *technosolutionism* (Atanasoski & Vora, 2019), which sees technological innovation alone as a possible means of solving social problems, and to develop concrete yardsticks that allow us to examine, from time to time, how relations of social dominance, such as sexism and racism, are reimagined, modified, or reinforced through the use of technologies.

For these reasons, this study underscores 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. The analysis has revealed how AI systems can perpetuate and amplify existing gender biases, often reflecting the societal contexts in which they are developed. 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. By linking the concept of gender performativity to AI ethics, this research provides a novel perspective that emphasizes the socially constructed nature of gender, and the role AI systems play in reinforcing these constructions. 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.

## References

Anders, G. (2002). *Die Antiquiertheit des Menschen. 1: Über die Seele im Zeitalter der zweiten industriellen Revolution* (2. Aufl. in der Beck'schen Reihe). Beck.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*.

Arendt, H., Allen, D. S., & Canovan, M. (2018). *The human condition* (Second edition). The University of Chicago Press.

Atanasoski, N., & Vora, K. (2019). *Surrogate humanity: Race, robots, and the politics of technological futures*. Duke University Press. https://doi.org/10.1215/9781478004455

Austin, J. L., & Urmson, J. O. (2009). *How to do things with words: The William James lectures delivered at Harvard University in 1955* (2. ed., [repr.]). Harvard Univ. Press.

Bachmann-Medick, D. (2016). Chapter II: The Performative Turn. In *Cultural Turns* (pp. 73–102). De Gruyter. https://doi.org/10.1515/9783110402988-004

Barad, K. M. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press. https://doi.org/10.2307/j.ctv12101zq

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press.

Behrent, M. C. (2013). Foucault and Technology. *History and Technology*, *29*(1), 54–104. https://doi.org/10.1080/07341512.2013.780351

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). 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*, *187*(4175), 398–404. https://doi.org/10.1126/science.187.4175.398

Bishop, J. M. (2020). *Artificial Intelligence is stupid and causal reasoning won't fix it* (Version 1). arXiv. https://doi.org/10.3389/fpsyg.2020.513474

Bjerring, J. C., & Busch, J. (2024). Artificial intelligence and identity: The rise of the statistical individual. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01877-4

Boddington, P. (2023). The Rise of AI Ethics. In P. Boddington, *AI Ethics* (pp. 35–89). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9382-4_2

Bourdieu, P. (2001). *Masculine domination* (1. publ). Polity Press.

Braidotti, R. (2013). *The posthuman*. Polity Press.

Broome, J. (1991). V—Fairness. *Proceedings of the Aristotelian Society*, *91*(1), 87–102. https://doi.org/10.1093/aristotelian/91.1.87

Browne, S. (2015). *Dark Matters: On the Surveillance of Blackness* (p. dup;9780822375302/1). Duke University Press. https://doi.org/10.1215/9780822375302

Bunge, M. (1963). *A General Black Box Theory*. *30*(4), 346–358. https://doi.org/10.1086/287954

Butler, J. (1997). *The psychic life of power: Theories in subjection*. Stanford University Press. https://doi.org/10.1515/9781503616295

Butler, J. (2004). *Undoing gender*. Routledge. https://doi.org/10.4324/9780203499627

Butler, J. (2006). *Gender trouble: Feminism and the subversion of identity*. Routledge.

Butler, J. (2011). *Bodies that matter: On the discursive limits of "sex."* Routledge. https://doi.org/10.4324/9780203828274

Butler, J., & Shusterman, R. (1999). *Performativity's social magic*.

Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, *22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

Cavarero, A. (2016). *Inclinations: A critique of rectitude*. Stanford university press.

Chang, X. (2023). Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm. *Advances in Economics, Management and Political Sciences*, *23*(1), 134–140. https://doi.org/10.54254/2754-1169/23/20230367

Chu, K. H., Brown, N. J., Pelecanos, A., & Brown, A. F. (2018). Simpson's paradox: A statistician's case study. *Emergency Medicine Australasia*, *30*(3), 431–433. https://doi.org/10.1111/1742-6723.12943

Clemons, T. R. (2014). Blind injustice: The Supreme Court, implicit racial bias, and the racial disparity in the criminal justice system. *Am. Crim. L. Rev.*, *51*, 689.

Coeckelbergh, M. (2024). *Why AI undermines democracy and what to do about it*. Polity Press.

Cornell, D., Rosenfeld, M., & Carlson, D. G. (Eds.). (2016). *Deconstruction and the Possibility of Justice*. Routledge. https://doi.org/10.4324/9781315539744

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. https://doi.org/10.12987/9780300252392

De Lauretis, T. (1984). *Alice doesn't: Feminism, semiotics, cinema*. Indiana University Press. https://doi.org/10.1007/978-1-349-17495-9

De Lauretis, T. (1987). *Technologies of gender: Essays on theory, film, and fiction*. Indiana University Press. https://doi.org/10.1007/978-1-349-19737-8

De Lauretis, T. (1990). Eccentric Subjects: Feminist Theory and Historical Consciousness. *Feminist Studies*, *16*(1), 115. https://doi.org/10.2307/3177959

De Lauretis, T. (1994). *The practice of love: Lesbian sexuality and perverse desire*. Indiana University Press.

Derrida, J. (2018). *Before the law: The complete text of Préjugés* (S. Van Reenen & J. De Ville, Trans.). University of Minnesota Press. https://doi.org/10.5749/j.ctv75d0d9

Dişci, Z. (2024). Ideology, Subject and Gender: Undoing Representations in the Thought of Teresa De Lauretis and Judith Butler. *Feminist Encounters: A Journal of Critical Studies in*

*Culture and Politics*, *8*(1), 20. https://doi.org/10.20897/femenc/14231

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818. https://doi.org/10.1111/isj.12370

Drage, E., & Frabetti, F. (2023a). AI that Matters: A Feminist Approach to the Study of Intelligent Machines. In J. Browne, S. Cave, E. Drage, & K. McInerney (Eds.), *Feminist AI* (1st ed., pp. 274–289). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192889898.003.0016

Drage, E., & Frabetti, F. (2023b). 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*, 016224392311646. https://doi.org/10.1177/01622439231164660

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor* (First Picador edition). Picador St. Martin's Press.

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Floridi, L. (2013). Technology's In-Betweeness. *Philosophy & Technology*, *26*(2), 111–115. https://doi.org/10.1007/s13347-013-0106-y

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001

Foucault, M. (2008). *The history of sexuality: The will to knowledge: vol. 1*. Penguin.

Gursoy, F., & Kakadiaris, I. A. (2022). Equal Confusion Fairness: Measuring Group-Based Disparities in Automated Decision Systems. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 137–146. https://doi.org/10.1109/ICDMW58026.2022.00027

Haraway, D. J. (2016). *Manifestly Haraway*. University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816650477.001.0001

Hayles, N. K. (2017). *Unthought: The power of the cognitive nonconscious*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226447919.001.0001

Hayles, N. K. (2022). Inside the Mind of an AI: Materiality and the Crisis of Representation. *New Literary History*, *54*(1), 635–666. https://doi.org/10.1353/nlh.2022.a898324

Hirschmann, N. J., & Regier, E. F. (2019). Mary Wollstonecraft, Social Constructivism, and the Idea of Freedom. *Politics & Gender*, *15*(4), 645–670. https://doi.org/10.1017/S1743923X18000491

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of

antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 513–513. https://doi.org/10.1145/3351095.3375674

Irigaray, L. (1985). *Speculum of the other woman*. Cornell University Press.

Jasanoff, P. S., Markle, D. G. E. E., Peterson, J. C. C., & Pinch, D. T. J. (2001). *Handbook of Science and Technology Studies*. SAGE Publications.

Jonas, H. (1984). *The imperative of responsibility: In search of an ethics for the technological age*. Univ. of Chicago Press.

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. https://doi.org/10.1145/3022181

Kohler-Hausmann, I. (2017). The Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3050650

Lagioia, F., Rovatti, R., & Sartor, G. (2023). Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI & SOCIETY*, *38*(2), 459–478. https://doi.org/10.1007/s00146-022-01441-y

Latour, B. (1993). *We have never been modern* (C. Porter, Trans.). Harvard University Press.

Licoppe, C. (2010). THE 'PERFORMATIVE TURN' IN SCIENCE AND TECHNOLOGY STUDIES: Towards a linguistic anthropology of 'technology in action.' *Journal of Cultural Economy*, *3*(2), 181–188. https://doi.org/10.1080/17530350.2010.494122

Lloyd, G. (1979). THE MAN OF REASON. *Metaphilosophy*, *10*(1), 18–37. https://doi.org/10.1111/j.1467-9973.1979.tb00062.x

Mbembe, A. (2020). *Brutalisme*. La Découverte. https://doi.org/10.3917/dec.mbemb.2020.01

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Northpoint. (2015). *Practitioner's Guide to COMPAS Core*. https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books.

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information* (First Harvard University Press paperback edition). Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pearl, J. (2003). CAUSALITY: MODELS, REASONING, AND INFERENCE, by Judea Pearl, Cambridge University Press, 2000. *Econometric Theory*, *19*(04). https://doi.org/10.1017/S0266466603004109

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (First edition). Basic Books.

Plato, & Derrida, J. (2004). *Phèdre* (L. Brisson & Plato, Eds.; Nouvelle édition corrigée et mise à jour). GF Flammarion.

Rawls, J. (2003). *Justice as fairness: A restatement* (E. Kelly, Ed.; 3. printing). Belknap Press of Harvard University Press.

Rode, J. A. (2011). A theoretical agenda for feminist HCI. *Interacting with Computers*, *23*(5), 393–400. https://doi.org/10.1016/j.intcom.2011.04.005

Rubin, G. (2012). The Traffic in Women: Notes on the "Political Economy" of Sex. In *Deviations* (pp. 33–65). Duke University Press. https://doi.org/10.1215/9780822394068-002

Ruggieri, S., Alvarez, J. M., Pugnana, A., State, L., & Turini, F. (2023). Can We Trust Fair-AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 15421–15430. https://doi.org/10.1609/aaai.v37i13.26798

Ryan, A. (2006). Fairness and Philosophy. *Social Research*, *73*(2), 597–606. https://doi.org/10.4159/harvard.9780674736061

Sedgwick, E. K. (1993). *Tendencies*. Duke Univ. Press. https://doi.org/10.1215/9780822381860

Shapiro, A. (2017). Reform predictive policing. *Nature*, *541*(7638), 458–460. https://doi.org/10.1038/541458a

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Siess, J. (2005). Un discours politique au féminin. Le projet d'Olympe de Gouges. *Mots*, *78*, 9–21. https://doi.org/10.4000/mots.293

Stiegler, B., & Ross, D. (2013). *What makes life worth living: On pharmacology* (English edition). Polity.

Vaccaro, M. A. (2019). *Algorithms in human decision-making: A case study with the COMPAS risk assessment software*.

Van Nood, R., & Yeomans, C. (2021). Fairness as Equal Concession: Critical Remarks on Fair AI. *Science and Engineering Ethics*, *27*(6), 73. https://doi.org/10.1007/s11948-021-00348-z

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (Vol. 11839, pp. 563–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51

Young, I. M. (1990). *Justice and the politics of difference* (Nachdr.). Princeton Univ. Press.

Ziosi, M., Watson, D., & Floridi, L. (2024). A Genealogical Approach to Algorithmic Bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4734082