

CAN XAI BE ETHICALLY JUSTIFIED?

How to rethink the concept of explainability in AI through feminist standpoint theory and marginalized groups' experiences

Marilou Niedda  m.niedda@uu.nl

Freudenthal Institute, Utrecht University, The Netherlands

Article type: Research article

Review process: Double-blind peer review

Topical collection: Ethics and Normativity of Explainable AI: Explainability as a Social Practice, guest editors: Tobias Matzner, Suzana Alpsancar, Martina Philippi, Wessel Reijers

This open-access article is published with a Creative Commons CC-BY 4.0  license

<https://creativecommons.org/licenses/by/4.0/>

DOI: <https://doi.org/10.59490/jhtr.2026.4.7956>

ISSN: 2773-2266

Submitted: 19 December 2024 **Revised:** 21 July 2025, 11 March 2026, 28 May 2026 **Accepted:** 16 June 2026 **Published:** 13 October 2026

How to cite (APA): Niedda, M. (2026). Can XAI be ethically justified? How to rethink the concept of explainability in AI through feminist standpoint theory and marginalized groups' experiences. *Journal of Human-Technology Relations*, 4, 1-19. <https://doi.org/10.59490/jhtr.2026.4.7956>

Corresponding author: Marilou Niedda

Keywords

XAI; Validity; Feminist
Standpoint Theory;
Propositional Knowledge;
Experiential Knowledges

Abstract

This article aims to reconcile the technical and epistemic dimensions of Explainable AI (XAI) with its ethical components. I argue that XAI's validity is comprised of its model application and verification from a technical viewpoint in AI design. However, this concept and methodology also serve ethical aims by examining algorithmic calculations a posteriori, which can produce computationally generated inequities towards marginalized groups. By linking its epistemic and ethical goals, I examine how XAI's models can be improved through a political, social, and epistemological framework centered on the knowledge production of discriminated groups. Feminist standpoint theory highlights the context in which knowers situate themselves and inform their cognitive contents with their social locations and experiences. Emphasizing social contexts in the case of XAI allows me to expand its technical validity to subjective experiences of oppression that are structurally sustained in the Global North. By doing this, I argue that rethinking the use of XAI would ultimately enhance the field of AI ethics.

1 INTRODUCTION

Explainable AI, shortened to XAI, is a central concept in the practices and research on AI technologies. It refers to the methods and models used by designers and engineers developing AI, highlighting the technical processes that yield specific outputs; the internal workings of an AI system are explained in the sense that its algorithmic calculations can be made traceable. As AI results are generally opaque, XAI is intended to make sense of what seems to be deeply technical and epistemologically obscure in AI technologies.

Ingrained epistemic debates discuss what a good, robust, or useful explanation in XAI constitutes, both in computer science and in philosophy (Curtis-Trudel, 2023; Khalifa, 2012; de Regt, 2012). The original meaning of XAI is grounded in its technical deployment; a good XAI model explains x through causal algorithmic reasoning. However, this seems not to be enough to provide a good explanation, one that makes sense socially and not only technically: people should be able to understand AI systems in their broader social context. Indeed, if designers of AI systems must render their algorithmic work explainable *a posteriori*, it is not solely for matters of epistemic accuracy, but also for ethical reasons; XAI must foster epistemic transparency (High-level expert group on AI, 2019), to prevent any harmful uses (and design) of such technologies that can negatively impact populations. Examples are unfortunately numerous, from the infamous COMPAS case in the US, where black inmates were more discriminated against than their white counterparts for risk of reoffending (Angwin et al., 2016), to the use of a sexist recruiting tool by Amazon (Dastin, 2018). From these cases emerges a right for the general public to understand the reasons why such technologies can perpetuate structural injustices towards vulnerable populations, but not exclusively. With this in mind, explanations cannot remain purely technical; they must emphasize the social contexts in which those AI systems operate, who uses them, where, and for what reasons. There is an urge among both researchers and practitioners to make XAI explanations more transparent and accessible, especially for broader audiences. In this work, I therefore focus on how AI practitioners can improve the accuracy of their explanations using XAI.

One may wonder, why then not work on the concept of interpretability instead of XAI? Interpretability in AI research is about providing a deeper understanding of the social context surrounding the use of an AI technology (McGrath & Jonker, 2024; Amodei, 2025). It might be more intuitive to work on this concept rather than risk increasing conceptual confusion and inflation surrounding a necessary social deployment of XAI. Explainability would be heuristically technical, whereas interpretability would be social. Nevertheless, XAI is a unique concept and method in AI research because it implies that there are different kinds of explanations, and the term admits a multitude of viewpoints on who, why, and how a given phenomenon deserves an explanation. As a matter of fact, explanations cannot be solely technical; XAI represents the linkage between an AI system's inner and outer workings, as technical highlights can be meaningless without social context. From considering XAI as such, I discuss the latter concept – rather than interpretability – because it is *ethically fruitful* for understanding the impact of AI technologies in our societies, in virtue of carrying norms about what constitutes a valid explanation.

In this work, I focus on the relationship between epistemic and ethical concerns within XAI methodology. To do so, I explore what kinds of socially-inclined knowledge XAI ought to highlight to provide more *useful explanations* of an AI system. Technical content is, as a matter of fact, not the only form of knowledge present in AI objects. As I wish to make it intrinsically ethical in its justification, I defend a repurposing of explainability as primarily a political and social method/concept. Here, I follow Sally Haslanger's (2000) view on the *ameliorative* necessity of concepts, where they should be epistemically useful by socially improving our conditions of existence. The analytical feature of concepts within theory-making addresses the

question of what the point of x is and what it enables us to accomplish politically and socially. Then, XAI should not only be technically descriptive but also socially beneficial. This epistemic revisitation has empirical utility for AI practitioners, as it pushes them to think more critically about the overall usage of their AI systems. With this contribution, I aim to advance this agenda by detailing the socially located questions I intend to incorporate into the XAI method.

To provide social considerations useful to the XAI methodology, I turn to feminist theory of knowledge, particularly standpoint theory. More precisely, I draw on the works of ‘feminist standpoint epistemologists’ (Frasch, 2019), such as Sandra Harding, Donna Haraway, or Alison Wylie, who are philosophers/historians of science. This branch of standpoint theory is, above all, concerned with the impact of social identity on the making of the sciences. The reasons why I do so are that (i) XAI needs to conceptually innovate by reinforcing bias detection to prevent harm to vulnerable populations, by focusing on social features such as gender or race. Moreover, (ii) I want to show that formal knowledge in XAI can be biased. Through (i) and (ii), I argue that there is a political responsibility in AI research that engineers have to endorse to improve explanations with XAI. Indeed, the field of AI is not exempt from a broader critique of its conceptual tools, particularly when AI systems can perpetuate or reinforce structural injustices. I will show that feminist standpoint theory expands the ethical justification of XAI, as it deconstructs the epistemological foundations and political practices that have been detrimental to the existence of marginalized groups (Grasswick, 2011). This article discusses how to include standpoint theory’s epistemic contributions to XAI methods. To do this, I draw notably on the work of Alison Adam (1998, 2000), which allows me to explain the gendered and androcentric character of early AI research. This motivates the reference to standpoint theory to better comprehend designs of AI-related methodologies, as feminist epistemology is primarily concerned with the impact of gender on knowledge formation. Additionally, I focus on a case study developed by Huang et al. (2022) in which they test a XAI model for reducing bias; I chose the latter for its comprehensiveness and the clear linkage the authors make between feminist and XAI methodologies.

To reconcile the epistemic-ethical features in XAI, my reasoning is divided into two parts, corresponding to the two main arguments I present. The first one is *epistemically descriptive*, as I discuss the epistemological tradition in which XAI is embedded: positivism. It corresponds to the realist view, which holds that reality is ‘unveiled’ through the constant discovery of natural elements. XAI’s epistemic grounding is mathematically oriented, and therefore fails to include social explanations of algorithmic use. I notably draw on the work of Alison Adam (1998, 2000) to support my view. I also argue in this part that the current application of ethical XAI is too normative to be empirically impactful for the most marginalized groups.

The second argument I make concerns the *meliorative* aim of standpoint theory: which social considerations should we include in XAI, and how can we provide rightful explanations for cases of structural discrimination? For that, I draw on (i) the situated knowledge and (ii) the epistemic advantage theses (Crasnow, 2014). Broadly, standpoint theory is about situating knowers’ social identities and sorting which knowledge is more useful and truer in certain social situations. This sorting benefits the largest number of people in the long run and can be normative within XAI models.

My article is conceptual in nature. I reinforce the view that concepts in AI research must be enhanced within social epistemology, and that XAI is ontologically valid if and only if it serves ethical purposes. I do not provide extensive empirical responses on how AI experts could step-by-step improve their practices. This would require another article on the matter, and I wish to provide conceptual grounding for it. However, I consider that highlighting epistemic-ethical tensions related to XAI offers fertile ground for further research.

2 A NECESSARY EPISTEMIC AND ETHICAL COUPLING

To explore the relationship between an epistemic and ethical application of XAI, I first outline that, albeit a technical concept primarily concerned with model validity in computer science, XAI is informed by ethical considerations. I discuss the epistemological tradition in which XAI is embedded (positivism), and then explore the current ethical approaches to the concept. I present the limitations of both epistemological/ethical views.

2.1 THE MODEL VALIDITY OF XAI AND THE OBJECTIFICATION OF REALITY

Firstly, I argue that XAI, as a method and concept, is technical because it is embedded in a specific epistemological tradition centered on formal language. Indeed, XAI's epistemic grounding is primarily mathematical and logical, which is traditionally situated within positivist methods.

In the 19th century, Auguste Comte generalized positivism, a school of thought that assumes that real knowledge is empirical and quantifiable through a limited set of methods (Feichtinger et al., 2018). To prove x is to gather evidence of the latter through experiments and measurements. The role of numbers, through the quantification of an x phenomenon, is pivotal because they incarnate natural laws that need to be uncovered to generalize truth-claims. There is therefore a clear distinction between the subject/object of knowledge, through the objectification of the social world (Harding, 1992). There seems to be no other way to express scientific discovery, as the use of formal language is more objective and therefore, more real. This equation matters because it grants epistemic authority to the concept of objectivity in the making of science.

According to this view, the most robust knowledge in the sciences must come from this tradition of objectified reality, where the translation of shaping knowledge (through experimentation or else) is formally transcribed. This often includes numerical evidence, in which reality can always be mathematically shaped, leading to the impression that objective knowledge is supported by numbers (Porter, 1995). Formalized knowledge is then perceived as the highest form of knowledge, through propositional transcription. Propositional knowledge can either be true or false, and can be verified through inference or perception. This means that well-formulated propositions must be true to be accepted, and all the more, formalized to not let any subjective elements tarnish their truth-value. Ideologically, this influences how people perceive artificial intelligences, as they are able to process huge amounts of numerical data through their algorithmic structure. Algorithms, in a positivist fashion, can be understood as neutral in virtue of being enacted by technological artefacts that simply follow numerical rules. It is therefore easy to assume that the explanations provided by XAI models target mere technical errors, and are not concerned with understanding complex social contexts.

Historically, good explanations in the sciences inherit from this tradition of securing evidence through numerical quantification (Porter, 1995). Then, there is little wonder that XAI (if explainability equates to explanation) is all about model validity. Engineers have to explain the deep internal workings of a machine, even if it makes little sense to many others, and sometimes even to them. XAI inherits from this tradition of knowledge production and is regarded as more objective when dealing with algorithmic internal workings, because numbers carry explanatory power¹ (Fel et al., 2022). The legitimacy of numbers in ML (machine learning) can (i) give the illusion that something is always true in virtue of being formalized, and (ii) provide partial explanations about the use of an AI by a given social actor. Propositional

¹ What I mean by *explanatory power* is the level of accuracy of an explanation. The more explanatory power x has, the more accurate the explanation is.

thinking, through formalization, wipes out social context in that regard by anchoring reality in an objectifying truth.

Drawing on this tradition of knowledge formation, XAI's explanations are mainly technical. Indeed, it is concerned first and foremost with a posteriori technical verification of existing algorithms, meaning that XAI models are not easily understandable to everyone. XAI's goal is to enhance an AI by generating targeted knowledge that improves ML systems through linear regressions, decision trees, or random forests (among others). Explanations are then local, internal, and model-based (Bellucci et al., 2021).

The issue with this concept often concerns articulating what constitutes a good explanation in the first place, and many papers on XAI aim to do so (Byrne, 2023). In its current underpinning, the most rightful explanations in XAI are technical. However, working on a normative definition of what *the best* explanation for XAI is can be epistemically confusing, as this dispute is not even settled in classic philosophy of science. Debates about explanations are quite contentious, to the point that many philosophers turn to the term understanding instead (Khalifa, 2012; de Regt, 2012). In practice, even (natural) scientists do not use the concept in the same way as theorists and philosophers, and the same goes for AI practitioners. For instance, the High-level expert group on AI (2019) of the European Commission showcases the explainability-explanation coupling in its "Ethics Guidelines for Trustworthy AI":

[P]rocesses [of AI systems] need to be transparent, the capabilities and purpose [...] openly communicated, and decisions [...] *explainable* [emphasis added] to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An *explanation* [emphasis added] as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. [p. 15]

The easiest way to materialize a *good* explanation is through the coupling of an *explicandum* (an object of study) illustrated by an *explicatum* (a concept used to simplify an object) (Carnap, 1950). In the case of XAI, the *explicatum* needs to be made transparent through its *explicandum*, as one needs to know both the reasons why an AI system was used in a certain way at a specific time and place (social explanation) and how its results were obtained (technical explanation).

My focus point in this work is not on explanation per se, as the concept remains highly contested within the philosophy of science in the first place. Instead, I chose to focus this paper on exploring the ethical potential of XAI models by assessing whether it is possible, and desirable, to incorporate non-formal dimensions into their core. Therefore, rather than advancing a conceptual examination of explanation, I investigate the *validity* of this concept in light of the aims it is supposed to serve. I indeed question whether XAI's overall validity is relevant when its models do not articulate the social context in which algorithms are created and/or used. Algorithmic calculations cannot, indeed, convey social norms because they are mathematically driven, but like every model, XAI attempts to represent a *real-world system*, which is first and foremost experiential. As Helen Longino (1993, p. 114) puts it, a "real world system [is] based on an analogy between the world and the system, that is the supposition that the model and the system share some significant features in common". With this in mind, one can conclude that XAI is not primarily about showcasing what constitutes a good explanation according to diverse audiences (Al-Ansari et al., 2024), but rather about the validity of a model in describing a real-world feature in an AI system. It is therefore important to open up XAI's methodology to social components.

Kristina Rolin (2006) explains that the use of social components is justifiable in epistemology within a contextualist theory of epistemic justification (p. 3), but not within a foundationalist

framework often assumed in mathematically oriented disciplines. Indeed, a foundationalist theory holds that assumptions are justified by unchallenged “basic beliefs” (p. 4), such as theorems. For instance, an algorithm’s authority is granted relative to its (objective) mathematical structuring – numbers are neutral – and its ability to process large datasets. Within this framework, it is possible to *integrate* phenomenological considerations into XAI’s methodology.

2.2 XAI AND ANDROCENTRISM

In this part, I argue that the positivist framework in which XAI is embedded provides technical explanations mainly offered by *male logicians* concerned with AI. This reduces the explanatory power of XAI and shows that human biases can be pronounced in formal language.

The feminist Science and Technology Studies (STS) researcher Alison Adam conducted in-depth research on the influence of mathematical and logical reasoning that underpins androcentric biases in AI-based computer systems. Adam (1998) articulates in “Artificial Knowing” that AI is an umbrella term from the 1960s designating “the science of making machines do things that would require intelligence if done by *men* [emphasis added]” (p. 34). Despite her study being written in 1998, I consider her analyses of gender bias incorporated into AI methods and concepts at the time still applicable today and to modern ML techniques. She indeed demonstrates that historically, the first AI machines were designed to embody (hu)man knowledge by making every epistemic step in a reasoning process clear. This idealistic view of knowledge production relies heavily on symbolic language and mathematics to elucidate such steps. For instance, she explains that the first AI technologies incorporate problem-solving methods inspired by chess, a game played by numerous white, male logicians at the origin of AI. This way of thinking assumes that knowers are autonomous, rational at all times, and can produce a perfectly traceable epistemic process that can be reproduced for problem-solving. Adam makes clear the association of mathematical methods with the figure of a male, white computer scientist, assuming that knowledge can be extracted from their social contexts, processed and solidified, which is the main assumption of positivism. Although her study focuses on case studies using symbolic AI, an approach that involves representing knowledge and reasoning through explicit symbols and logical rules, I argue that her research can be applied to ML techniques, and therefore, to XAI.

In fact, the distinction between “symbolic AI” and “connectionist models” (Adam, 1998, p. 41) – or neural networks in ML – does not assume the same design of *reasoning* in AI. Contrary to symbolic AI, ML patterns are extracted from data and replicated, forming a brain-like structure. Because of this, rather than relying on the expert knowledge present in symbolic AI, ML techniques are much more unpredictable, as there are no systematic classifications of data that follow the same pattern. That is one of the reasons that making an ML model ‘bigger’ in terms of coverage is riskier, as data differ substantially. The concept of XAI was developed to comprehend certain ML predictions a posteriori, thereby expanding a model’s applicability by making its epistemic process clear. Despite that, and even if ML relies less on expert knowledge than in symbolic AI, the same techniques are implemented, such as logical rules, mathematical thinking, or statistical reasoning. Similar epistemic assumptions about mathematical knowledge (autonomous, uncontextualized forms of reasoning) are present within neural networks, assuming that androcentric, Western-oriented biases are incorporated into the collection and processing of data. This model relates to the justified true beliefs (JTB) approach, in which AI practitioners focus on articulating logical, true thinking within AI systems. This is pivotal in the symbolic approach and essential for ML models. However, if the goal of XAI is to offer the *most neutral* explanations of algorithms, social biases ought to be detected. I consider that the formal core of XAI does not allow for a crucial investigation of social contexts.

2.3 THE ETHICAL IMPORTANCE OF XAI

Given that focusing on technical aspects within XAI models does not provide sufficient explanations, as AI artefacts can carry out biases, this concept also integrates ethical concerns.

Al-Ansari et al. (2024), in their systematic literature review on “Explainable AI”, state that “XAI seeks to develop machine learning (ML) techniques that produce explainable models while not interfering with the learning performance of AI techniques used to evaluate data” (p. 1). The epistemic goal of such ML techniques is to increase human trust in AI systems in the foreseeable future, whilst improving their predictive performance. Indeed, XAI models are applicable in pivotal domains in public policy, including medicine, finance, and transportation (p. 1), and now constitute a central principle in the 2024 EU Artificial Intelligence Act (referred to as AIA; see the 2023 short report from the Technology and Privacy Unit of the European Data Protection Supervisor (2023)). All of this shows the political importance of XAI and its applicability to intentionally create more ethical AI systems, where humans must fully comprehend why a particular algorithmic process leads to a given outcome. Therefore, XAI cannot be valid solely by remaining technically focused. For instance, its methods ensure that engineers provide documentation related to an algorithm, making the data used traceable.

As a matter of fact, the most contentious issue with current ML techniques is that many algorithmic results can only be understood a posteriori (Fazi, 2020), despite serving the intended goals of a given AI system efficiently: the algorithm can work, whilst not being able to explain itself. This aspect is challenging for practitioners who cannot fathom, predict, or explain why a given AI system reached a particular conclusion. Neural networks, characterized by ML methods, blur the distinction between the inputs and outputs of AI systems, resulting in the infamous black box problem². Consequently, a black-boxed AI leads to a lack of transparency in the causal explanations of algorithms (Vaassen, 2022), highlighting the need for epistemic transparency to reduce opacity. Indeed, XAI’s role is also to ensure that AI systems are more ethical by being more *transparent*; the term Ethical AI refers to systems ensuring diverse and fair outcomes to “avoid biases that can cause harm and discrimination. In real-life examples, AI technologies have infringed on [...] attributes such as gender, race, and ableism” (Zowghi & Bano, 2024, p. 1). By opening the black box, AI experts can retrospectively explain why a particular algorithm may have been detrimental to targeted populations. Cases of algorithmic bias are prime examples in both the public and private sectors; they reflect computationally generated inequities. We refer to this concept to designate the (re)creation of *structural discriminations* (such as of racist, or racist nature) by *artificial agents* (such as algorithms, performed by computers). For instance, if an AI recruiting tool had been trained only on a dataset of white male CVs, there is a high chance it would discriminate against other groups, such as women (Dastin, 2018). Algorithms do not create existing discrimination *per se* (as they are already ingrained in our social norms and political systems), but rather reinforce it. In that sense, XAI methods can help uncover undesirable discrimination that certain AI systems (re)create.

Given the ethical role of XAI, I formulate XAI’s *epistemic core* as follows:

P1: XAI’s explanations are embedded in its model representation of algorithmic calculations.

P2: XAI’s explanations extend to ethical concerns.

P3: Explanations in XAI must be both technical and social.

C: XAI is valid when it serves epistemic-ethical goals simultaneously.

² See also citation in [sub-section 2.1](#).

2.4 THE LIMITS OF XAI AS AN ETHICAL CONCEPT

Given the limited character of formalized explanations, XAI is also about taking into account ethical considerations in AI systems. However, I consider that the current ethics sustaining XAI as a concept/method is too normative to be empirically effective.

XAI concerns rendering algorithmic calculations epistemically transparent within ML techniques. However, AI practitioners may struggle to link their epistemic practices with ethical considerations, which are always context-dependent, as XAI methods are reductionist in nature. One can argue that this thin understanding of social context is not easy to achieve, especially for people working with technical skills. It requires more reflexivity and time than individuals trained in ethics who work in the AI field, on top of having technical expertise. I consider that the current application of XAI within European law, for instance, does not advance the conversation, as it remains highly normative and top-down. XAI is about transparency, interpretability, and fairness, but there is an evident struggle to understand what this means in practice. In the EU AI Act(AIA) (2023), companies must ensure that their ML systems are explainable, meaning fair, but these principles remain analytically weak if they seem empirically inapplicable. Instead of pushing for more social XAI, normative notions, which are not critically enhanced through particular contexts, are utterly limited and do not really indicate what engineers ought to do (nor even non-technical experts).

To elaborate on this point, the distinction between epistemic and ethical values is not new in the making of science. It is indeed common to consider that scientific activity must be rational and objective by overlooking ethical aspects that are unnecessary to epistemic processes (Code, 1993). Therefore, the development of ethical principles is not epistemically bad *per se*, but I consider that ethics must be embedded and taught in context, alongside epistemic practices. Consequently, this normative account of ethics makes the divide between social/technical XAI even bigger.

Because of this unclear implementation of ethical XAI in the existing literature on the matter, Biecek and Samek (2024) argue that the ethical component of XAI should ultimately be put to the side. For them, the validation-oriented feature (model-like) of XAI is supported by the principles of “Research, Explore, Debug” (RED XAI), and the value-oriented one (ethical-like) is shaped by “Trust, Responsibility, Legality and Ethics” (BLUE XAI) (p. 1). They chose this distinction in reference to the RED/BLUE teams in IT security systems, referring to the XAI method for securing data. For them, and because of the ethical development of XAI, the concept is “experiencing a crisis of identity” (p. 1), in which the RED and BLUE goals are blurred altogether. This results in distinct communities of researchers working on this, leading to ethical XAI overshadowing epistemic XAI. They conclude with suggestions for streamlining various XAI models to develop more robust, systematic explanations of ML techniques.

They do so as they wish not to conflate epistemological-ethical aspects of XAI for two reasons; the first is methodological. Indeed, XAI for them is all about formalization, because its primary goal is to technically *explain* algorithmic calculations. Secondly, they argue that the literature on ethical XAI is not sufficiently precise. They consider that XAI is not a normative concept that can be used interchangeably with transparency or fairness in AI research; it serves specific epistemic purposes primarily addressed to AI practitioners with technical skills.

I share similar concerns about the highly normative character of concepts in most ethical AI applications, and XAI is not exempt from that. However, confining XAI to technical-only applications is problematic, as it fails to showcase the ethical commitments it is meant to embody. Explanations cannot only be technical in that regard, because most of the elements that are being transcribed in algorithms are coming from people’s experiences. Data are reductions of individuals’ information, and, as a result, XAI also serves a social role, highlighting certain elements in people’s lives to fully fulfil its explanatory role.

To Adam (1998), the ontological distancing between formal language and human experience within the AI sector has substantial repercussions on ethical considerations in AI epistemology. She advances that by “emphasising AI’s adherence to traditional epistemology, I am also arguing that it follows traditional moral theory, [so] collective moral responsibility is not a view predominantly found in AI systems” (1998, p. 80). Here, Adam implies that a positivist epistemological framework encourages the development of more individualistic and deontological approaches to ethics in AI. In this view, ethical considerations can be added a posteriori to AI concepts and practice (such as with XAI), and are less embedded in given social situations. There is indeed a broader debate on how ethical considerations can be articulated directly into the AI development process, resulting in an epistemic-ethical integrated approach (Russo et al., 2023). In a similar vein, Babushkina and Votsis (2022) argue that professionals using AI – in their case, health professionals – must understand the epistemic limits of machine-learning techniques to encourage better human decision-making. Therefore, a constant epistemic and ethical discussion has to simultaneously happen when certain AI technologies are involved. To conclude this part, XAI is embedded in a normative ethical tradition that does not provide concrete recommendations for how to use this concept to serve ethical goals. To do so, we need to understand XAI in relation to people’s actual experiences of discrimination by algorithms.

3 XAI THROUGH FEMINIST STANDPOINT THEORY

Enhancing XAI’s validity means rethinking how a technical concept can incorporate people’s real-life experience into its methodology. Indeed, AI practitioners bear epistemic authority regarding XAI, because they have sufficient technical knowledge to assess its model-like validity. This does not mean that (i) they fully understand the workings of an algorithm, especially if the social context around it is not well articulated, and (ii) that they remain the only ones able to decide what’s a good explanation of ML systems.

By people’s real-life experience, I designate the social context surrounding the use of certain data in an algorithm tested by an XAI model. In that sense, we need to target which information is epistemically useful to incorporate in an XAI model, in order to also serve its ethical purpose. XAI should not simply comprehend a posteriori artificial decision-making, but also provide a posteriori explanations about how a specific algorithmic calculation was or could be problematic to targeted groups of people, by reducing overall potential harm. What is epistemically ethical in XAI is to shed light on power relations within societies, aiming to improve overall knowledge about AI technologies.

I consider that the way knowledge is currently articulated by AI practitioners, and how the tech industry focuses on efficiency rather than the greater impact of their technologies on marginalized groups (Browne et al., 2023), limits XAI to formal, technical explanations. This epistemic practice embodies an epistemology of dualism, fundamental to positivist methods, sorting out good, useful, objective knowledge (for scientific enterprise) and bad, subjective knowledge (Hoquet, 2011); the latter has little to do with true empirical assessment of the natural world. Ethical principles in AI fall short because they do not address more peculiar, subjective experiences. And at the same time, relying on subjective experiences can be considered less objective and epistemically inferior for improving XAI models.

Those aspects motivate me to question the utility and applicability of XAI by enhancing its ethical justification through a context-sensitive epistemology that takes into account subjective experiences in scientific processes. Rather than confining XAI to technical explanations, we must collectively be ethically responsible by reconsidering which types of social explanations should be considered for this concept/method.

3.1 XAI AND EXPERIENTIAL KNOWLEDGE(S)

A fruitful component to understand data for an algorithm in context entails understanding which experiential knowledge(s)³ can be assessed by an XAI method, and involves some individual/collective reflexivity on the shaping of algorithms.

For Adam (2000), knowers in AI struggle to interpret and incorporate context-sensitive information into their mathematically oriented models – and couple the technical/ethical features of XAI – for the simple reason that the ongoing epistemology of AI is rooted in “traditional epistemology” (p. 97) and propositional knowledge, still within the positivist tradition. In this account of knowledge production, a piece of knowledge is true in virtue of *S* (an independent subject) knows that *p* (a proposition). Typically, propositional knowledge is well illustrated with examples in the natural sciences, which are less applicable to real-life events such as phenomenological experiences.

Thus, practical or experiential forms of knowledge have not been considered in classic epistemology as true knowledge by not fitting into the propositional model (Dalmiya & Alcoff, 1993). The knowing subject, in virtue of being *S* (autonomous and above any contextual elements defining them), must be able to formulate higher-order, logical, justified reasonings to be considered knowledgeable. Adam (1998) demonstrates that mathematicians, statisticians, and others are typically AI experts who adopt propositional forms of knowledge by default. For instance, data collected for ML models represent real-life situations that are reduced for processing. Converting experiential knowledge(s) into XAI models indeed presents an important methodological obstacle, which I will discuss further in this article (Adam, 2000).

When I refer to experiential knowledge(s), I draw on the work of Vrinda Dalmiya and Linda Alcoff (1993). They indeed show that varied knowledge(s) do not follow the propositional schema *S* knows that *p*, as they can also be experiential and practical. I already demonstrated that knowledge(s) processed and produced by AI systems follow this articulation of knowledge production through formalization. By contrast, experiential knowledge is often disregarded as *proper knowledge* given its embodied, social nature; “[the] schema ‘*S* knows that *p*’ is assumed to be adequate for all possible knowledge and as a consequence of this assumption, all knowing becomes propositional.” (p. 222). Within experiential knowledge(s), *p* can indeed be uncertain or less straightforward as it is built on specific know-hows that could be perceived as more intuitive, and not therefore justified according to classic epistemic standards. I consider that taking into account people’s social features in XAI methods (namely, experiential knowledge(s)) works for the betterment of XAI’s justification.

3.2 INTRODUCTION OF STANDPOINT THEORY

To emphasize the importance of highlighting people’s experiences within XAI methods, I argue that feminist standpoint theory is a valuable epistemic framework to do so. It is a social, practice-based epistemological approach that addresses social inequities by valuing the positionality of marginalized groups and encourages new grounds for truth(s) in the sciences. It criticizes the traditional canon in epistemology through the linkage of ethical considerations with epistemic findings. This is an important point, as XAI should be continually articulated in light of co-constitutive epistemic/ethical considerations.

Sandra Harding (2004) considers that standpoint theory constitutes a methodology with normative implications that encourages scientists’ social reflexivity to be developed in order to improve science-making, resulting in the “strong reflexivity” (Crasnow, 2014, p. 150). Strong reflexivity implies that knowers ought to reflect upon the processes through which they acquire

³ I put this plural because there is a multiplicity of knowing, which cannot encompass the way that knowledge cannot be a single enterprise.

knowledge, especially in science-making. Indeed, feminist epistemologists of science have long advocated for the intertwined character of epistemological attitudes with social ones; as we saw with Adam's work (1998), social attitudes influence scientific inquiry even in the most mathematically-driven disciplines (Harding, 1986; Longino, 1993). Engineers working with AI are not exempt from this, and rather than sorting out how to implement ad hoc ethical norms in their work a posteriori, they must accept that knowledge within XAI is also experiential.

On top of being epistemically reflexive, knowers' knowledge(s) are confined to one's "social location [which] systematically shapes and limits what [they] know, including tacit, experiential knowledge as well as explicit understanding, what [they] take knowledge to be as well as specific epistemic content" (Wylie, 2004, p. 343). Knowers are thus limited by their social and cultural surroundings, resulting in the constraint of epistemic production by these very limitations.

I claim that XAI would serve its ethical role in understanding algorithms by incorporating the social considerations of standpoint theory into its method. Two components in standpoint theory can be of use within an XAI methodology: the *situated knowledge(s) thesis*, which is about values in knowledge formation, and the *epistemic advantage thesis*, an epistemic stance acknowledging that "more accurate knowledge is likely to be generated from marginal social positions" (Sweet, 2020, p. 925; Crasnow, 2014). The third component of standpoint theory is about *political achievement*, as the adoption of such epistemological stances encourages marginalized groups' demands to be heard, seen, and politically recognized. I will work in more detail with the first two theses to exemplify XAI's concerns about ethics by considering people's subjective experiences in context. The political aspect of standpoint theory also matters in a broader political critique of the tech sector, but this is not the main focus of my work here.

4 BIASES' DETECTION THROUGH STANDPOINT THEORY: A CASE STUDY

In order to articulate the usefulness of standpoint theory for rethinking the social use of XAI by AI practitioners, I draw on a case study developed by Huang et al. (2022). In their work, they start their argument by stating that XAI is "techno-solutionist" (p. 3) in its methods and goals, as it does not sufficiently emphasize the structural character of bias detection, and they consider that the main role of XAI is to identify algorithmic biases after use. To substantiate their claim, they empirically test their hypothesis by running a program with the SHAP model (an XAI model) to highlight biases in a (fake) recruiting software. Fed with a couple of CVs, they ran the program and indicated a couple of variables' preferences, which emphasize the ability of applicants to work night shifts and have some past experience in the domain they are applying for. After running the model, the authors found that nothing was particularly wrong with the a posteriori classification of information based on the two sets of skills (night shifts and past experience). However, all the CVs that were rejected or ranked lower were those of black women. One reason is that some women are unable to work at night due to taking care of their children, and they also lack professional experience (due to potential past discrimination). They conclude that if the XAI model does not explicitly focus on 'identity-based' features such as gender or race, it is less efficient at detecting bias.

The issue in this case does not necessarily lie in the fact that black women are blatantly targeted, but rather that their identity is marked by proxy through other variables (such as their inability to work night shifts). Therefore, the XAI model in use is unable to identify this limitation, as it does not analyze structural power relations, despite the preferences of this recruiting software being technically valid.

Huang et al. (2022) state that the XAI model in question “assumes that to construct a general-purpose, objective observational tool, [it] only needs a relatively small knowledge base” (p. 5). This “small knowledge base” refers to technically clear instructions (with a preference for the ability to work night shifts and past experience) on how to achieve greater efficiency for a recruiter. The XAI model cannot see that a specific part of the population is being discriminated against, because it does not carry intentionality and meaning on its own, resulting in the lack of implementation of bias-reduction procedures. In this case, variables’ preferences are completely devoid of proxy definitions that could be articulated with a better grasp of oppressive social structures. Huang et al. then argue that “after all, gender and race are influencing hiring decisions, no matter how small the impact is” (p. 8) and wonder if “detecting and removing subtle biases like these [are] the main motivations for XAI” (p. 8). Then, the *raison d’être* of XAI is first and foremost bias detection around identity-related features for a posteriori verification.

Here, this case highlights a social usage of XAI that is not confined to a posteriori technical verification; it is about ensuring that no harm is being done to specific populations.

5 THE MELIORATIVE INFLUENCE OF THE USE OF STANDPOINT THEORY ON XAI

Drawing on the case study by Huang et al. (2022), I now discuss the contributions of standpoint theory to XAI. I examine the first two theses more closely and explore what they may entail for XAI’s methods.

5.1 SITUATED KNOWLEDGE(S)

One of the key components of XAI is the AI engineers with technical knowledge. To socially situate their viewpoints is therefore primary to *better understand* how an XAI model can work, by inferring its validity in relation to designers’ social situatedness in the epistemic realm. Donna Haraway (1988) asserts that knowledge claims are physically embedded in and situated within specific contexts, which limit the epistemic abilities of knowers and resist a universal reading of knowledge formation. Computational science is not exempt from this, in the sense that AI designers are always situated and shape models’ preferences. In fact, when AI practitioners are asked to provide an explanation of an algorithm, two things can be done through the lens of standpoint theory’s first thesis.

Firstly, many XAI models are considered *model-agnostic*, meaning they apply to different machine learning models (the SHAP model is one of them). Usually, those models are used when there is little access to the algorithm’s internal workings. The general issue with those models is that they assume that all algorithms work similarly, and that engineers have to overlook the data. Indeed, there is no possibility to trace the data being used if there is limited access to it. It is then easy to extrapolate that an algorithm is safe a posteriori when there is a universalization of criteria for verification. Therefore, as knowledge(s) are always contextual, it seems methodologically limiting to use a model-agnostic XAI without further investigation. AI practitioners should write down in detail why the model-agnostic XAI is limited in a given case, and maybe afterwards, refuse to grant the stamp ‘explainable’ to the stakeholder who wanted their algorithm to be explained. Furthermore, the type of data used must be traceable to ensure that explanations are complete. If that is not the case, explanations should be presented as partial in their concluding findings.

Secondly, AI designers have to assess their social positionality to reflexively interrogate an XAI model’s limitations. This can be materialized through self-inquiry, by wondering for whom/for what, by whom, the tested algorithm in question is being used. Those questions can be part of a

read-me guide within an XAI model and have to be answered in order to make the verification effective.

I am not stating that this would provide flawless explanations for any other stakeholders afterwards, but rather that it is better to have some of this information than none. This helps to socially contextualize the algorithms in use, and to emphasize the existence of the people who exist behind the data (Amoore, 2020).

5.2 THE EPISTEMIC ADVANTAGE THESIS

Standpoint epistemologists indeed claim that the least privileged groups offer “less partial and distorted accounts” (Wylie, 2004, p. 128) of social reality, as they do not hold privileged positions that they must maintain; their understandings of social situations are less mediated by their need to preserve specific power relations. Also, certain individuals have access to truer forms of knowledge – such as workers who are in direct contact with the objects they work with/on (Hartsock, 2003) – in virtue of having first-hand information.

Building on this encouragement of reflexive behaviours, I argue that diverse social perspectives must be taken into account to improve XAI. A crucial aspect of standpoint theory is that more accurate knowledge is likely to be generated from marginal social positions, which is an essentializing stance. In this account, it seems that being a black person grants you an automatic epistemic advantage over someone who is white, which is obviously problematic. There is more nuance to this argument, as epistemic advantage is not granted automatically; it involves some reflexive process about one’s condition, as being part of a social category is not enough (Crasnow, 2014). What is epistemically useful with a certain group affiliation is one’s ability to critically reflect upon the structural norms surrounding the shaping of their own existence. In a sense, one’s standpoint is both naturalist and constructivist, as it deconstructs pseudo-natural categories of identity formation whilst being necessary to fight structural oppression (Fuss, 1989).

How can this be helpful for the practical development of XAI methods? The ‘epistemic advantage’ thesis firstly implies that AI designers are to be from underprivileged groups to potentially improve XAI’s usefulness. Let’s imagine a group of AI engineers who are also particularly aware of (broad) inequities, and who can build ‘better’ XAI models. In certain cases, this can be effective⁴. However, there is an important limitation with this inclination. Indeed, companies like Google or Amazon excel at ethics washing by hiring people from diverse backgrounds who actually have little to say about the global impact of the technologies they work on (Harris, 2023). Small companies also follow their agendas and hire AI engineers to implement their vision. De facto, XAI as a method will not necessarily be improved by having people from underprivileged backgrounds who implement an entire research program, although there are promising feminist initiatives (State & Fahimi, 2023).

Rather than focusing on the identity of the engineers developing the XAI model, it is useful to formulate clear aims and goals about not discriminating against certain populations. For instance, in the Huang et al. case (2022), it would have been possible to emphasize socially related features (such as gender or race) alongside the ability to work night shifts. Some statistical inferences can be made in order to check the balance between people of color, women, and other applicants. Admitting the importance of those features matters, as XAI models can structurally address discrimination in diverse domains, such as the job market.

An important objection to this point is the risk of essentializing people’s identities when variables such as race or gender are treated epistemically as fixed. Being reduced to a social

⁴ Those cases are often from the private sector. For instance, collectives such as [The Data for Black Lives](#) project work exclusively with people of color to increase the representation of PoC people in the AI sector.

category can be extremely harmful, affecting collective and self-perception. However, I consider that reifying one's identity and personal experience into their gender or race is not the ultimate aim of designers' reflexive practices around XAI, as it defeats the purpose of reducing cases of AI-generated discrimination. Indeed, those socially constituted categories should be constantly critically assessed, but in the case of data, everything is reduced to very precise information. I therefore consider that it is the least damaging thing that one can do to always take into account the features of a given social group. At least, AI engineers can provide more detailed explanations of the populations represented in an algorithm, reinforcing acts of transparency about the data used. Explanations about this can remain consultative, but at least they should exist.

In this part, and through standpoint theory's main theoretical contributions, I showed that highlighting someone's social features – both in the case of AI engineers, and for the people represented in data points – helps reinforce XAI's explanatory force. Better explanations can be provided by emphasizing the social contexts in which people are embedded.

6 CONCLUSION

Sally Haslanger (2000) encourages us to enquire: "What kind of work should concepts do for us?", and "Why do we need them at all, and if we do need them, how can they be articulated to serve specific epistemic and social purposes?". This performative comprehension of conceptual work has been a key component of what I aim to explore in this paper through XAI. What is XAI supposed to achieve, if not solely computational validity? Is it still justified if we infer ethical cogitations to it, or does it become another normative concept in AI ethics? XAI appears to suffer from theoretical conflation to avoid further conceptual inflation in the field. With this in mind, my overall goal is to highlight the political and social role of XAI in reducing inequities produced by ML techniques.

To address this methodological and conceptual expansion, I first demonstrated in section 2 the current use of XAI from both technical and ethical perspectives. I pointed out the limitations of both approaches by articulating the technical character of XAI and its normative-ethical feature. I showcased that this concept inherits a tradition of knowledge production that considers formalized knowledge as the highest form of its kind. According to Adam (1998), the use of mathematics was often coupled with male characteristics, rendering the field of ML extremely masculine-oriented nowadays. In section 3, I demonstrated that standpoint epistemology is crucial to understanding phenomenological experiences to improve XAI. One must then take into account experiential knowledge(s) to better provide contextualized explanations. Engineers' positionality must be highlighted, and XAI explanations must focus on people's social features, such as race or gender. Section 4 deals with the possibility of XAI being used primarily as an ethical-epistemological concept/method to address and reduce biases in algorithmic systems. For that, I drew upon a case study showcased in the Huang et al. (2022) paper. To conclude, I examined in my last section how the first two theses of standpoint theory can positively contribute to the rethinking of XAI. AI engineers must (i) rethink the supposed value-neutrality of their work, and (ii) consider the importance of considering people's social features (race, gender, etc.) not as simple data points.

Quite broadly, AI research in philosophy is often viewed through the lens of more analytic frameworks. Having more work drawing from social epistemological approaches seems to be beneficial to the field, especially when one needs to understand the social dynamics at play between different (artificial) agents in such a booming discipline.

Data Access Statement

N/A

Contributor Statement

Marilou Niedda is the sole author of the article. She conducted all the collection of empirical data, as well as all of the curation, conceptualization, and writing.

Use of AI

During the preparation of this work, the author used GRAMMARLY in order to correct spelling mistakes and the overall level of English. After using this tool/service, the author(s) reviewed, edited, made the content their own and validated the outcome as needed, and take(s) full responsibility for the content of the publication.

Funding Statement

The author did not receive any financial support for the work on this article.

Acknowledgments

I want to heartily thank Karin Jongsma, Abigail Nieves Delgado, and Federica Russo for taking the time to provide extensive and numerous pieces of feedback on rewriting this article. I also thank the participants in Dina Babushkina's online seminar on the ethics and epistemology of AI for their expertise on this paper in January 2025. Finally, I thank the two reviewers who proofread the manuscript multiple times. Your comments have been very precious.

Conflict of Interest Statement

There is no conflict of interest.

References

- Adam, A. (1998). *Artificial knowing: Gender and the thinking machine*. Routledge.
- Adam, A. (2000). Deleting the subject: A feminist reading of epistemology in artificial intelligence. *Minds and Machines*, 10(2), 231–253. <https://doi.org/10.1023/a:1008306015799>
- Al-Ansari, N., Al-Thani, D., & Al-Mansoori, R. S. (2024). User-centered evaluation of explainable artificial intelligence (XAI): A systematic literature review. *Human Behavior and Emerging Technologies*, 2024(1), Article 4628855. <https://doi.org/10.1155/2024/4628855>
- Amodei, D. (2025, April). The urgency of interpretability. *Dario Amodei*. <https://www.darioamodei.com/post/the-urgency-of-interpretability>
- Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press. <https://doi.org/10.1215/9781478009276>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science*, 192, 241–250. <https://doi.org/10.1016/j.procs.2021.08.025>

- Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. *Ethics and Information Technology*, 24(2), 22. <https://doi.org/10.1007/s10676-022-09629-y>
- Byrne, R. M. J. (2023). Good explanations in explainable artificial intelligence (XAI): Evidence from human explanatory reasoning. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence Survey Track*, 6536-6544. <https://doi.org/10.24963/ijcai.2023/733>
- Biecek, P., & Samek, W. (2024). *Position: Explain to question not to justify*. <https://doi.org/10.48550/arxiv.2402.13914>
- Browne, J., Cave, S., Drage, E., & McInerney, K. (Eds.). (2023). *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines*. Oxford University Press. <https://doi.org/10.1093/oso/9780192889898.001.0001>
- Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.
- Code, L. (1993). Taking subjectivity into account. In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies*. Routledge.
- Curtis-Trudel, A. (2023). Mathematical explanation in computer science. *Philosophy of Science*, 90(5), 1079-1088. <https://doi.org/10.1017/psa.2023.26>
- Crasnow, S. (2014). Feminist standpoint theory. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of Social Science: A New Introduction*. Oxford University Press.
- Dalmiya, V., & Alcoff, L. (1993). Are “old wives’ tales” justified?. In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies*. Routledge.
- Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- de Regt, H. W. (2012). Understanding and explanation: Living apart together?. *Studies and History in the Philosophy of Science Part A*, 44(3), 505-509. <https://doi.org/10.1016/j.shpsa.2012.12.002>
- European Data Protection Supervisor. (2023). TechDispatch: Explainable artificial intelligence (#2/2023). Publications Office of the European Union. <https://doi.org/10.2804/802043>
- EU AI Act: First regulation on artificial intelligence. (2023, August 6). Topics | European Parliament. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- State, L., & Fahimi, M. (2023). Careful explanations: A feminist perspective on XAI. In *EWAF’23: European Workshop on Algorithmic Fairness*.
- Fazi, M. B. (2020). Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society*, 38(7–8), 55–77. <https://doi.org/10.1177/0263276420966386>

- Fel, T., Vigouroux, D., Cadène, R., & Serre, T. (2022). How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks. *Proceedings of the IEEE/CVF Winter Conference on Applications on Computer Vision (WACV)*, 720-730.
- Frasch, D. (2019). *Sujets de connaissance dans les théories du standpoint* [Knowing Subjects in Standpoint Theories] [Unpublished master's thesis]. ENS de Lyon.
- Feichtinger, J., Fillafer, F. L., & Surman, J. (Eds.) (2018). *The worlds of positivism: A global intellectual history, 1770 - 1930*. Palgrave Macmillan Cham. <https://doi.org/10.1007/978-3-319-65762-2>
- Fuss, D. (1989). *Essentially speaking: Feminism, nature and difference*. Routledge.
- Grasswick, H. E. (2011). Liberatory epistemology and the sharing of knowledge: Querying the norms. In H. E. Grasswick (Ed.), *Feminist epistemology and philosophy of science: Power in knowledge* (pp. 241–262). Springer. https://doi.org/10.1007/978-1-4020-6835-5_12
- Haigh, T. (2023, June 13). *AI: The Brand That Wouldn't Die* [Video]. Aachen University website. <https://khk.rwth-aachen.de/event/artificial-intelligence-the-brand-that-wouldnt-die-a-lecture-by-professor-thomas-haigh/>
- Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3), 575-599. <https://doi.org/10.2307/3178066>
- Harding, S. G. (1986). *The science question in feminism*. Cornell University Press.
- Harding, S. (1992). Rethinking standpoint epistemology: What is “strong objectivity”? *The Centennial Review*, 36(3), 437–470.
- Harding, S. (2004). Introduction: Standpoint theory as a site of political, philosophic and scientific debate. In S. Harding (Eds.), *The feminist standpoint theory reader: Intellectual and political controversies*. Routledge.
- Harris, J. (2023, May 22). “There was all sorts of toxic behaviour”: Timnit Gebru on her sack by Google. AI's dangers and big tech's biases. The Guardian. <https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases>
- Hartsock, N. C. M. (2003). The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In S. Harding & M. B. Hintikka (Eds.), *Synthese Library: Vol. 161. Discovering reality: Feminist perspectives on epistemology, metaphysics, methodology and the philosophy of science* (pp. 283–310). Springer. https://doi.org/10.1007/978-94-010-0101-4_15
- Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be?. *Noûs*, 34(1), 31–55. <https://doi.org/10.1111/0029-4624.00201>
- High-level expert group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

- Hoquet, T. (2011). *Cyborg philosophie: Penser contre les dualismes* [Cyborg Philosophy : To Think Against Dualisms]. Éditions du Seuil.
- Huang, L. T.-L., Chen, H.-Y., Lin, Y.-T., Huang, T.-R., & Hung, T.-W. (2022). Ameliorating algorithmic bias, or why explainable AI needs feminist philosophy. *Feminist Philosophy Quarterly*, 8(3/4). <https://doi.org/10.5206/fpq/2022.3/4.14347>
- Khalifa, K. (2012). Inaugurating understanding or repackaging explanation?. *Philosophy of Science*, 79(1), 15–37. <https://doi.org/10.1086/663235>
- Mackie, M. (1992). Review of *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, by H. E. Longino]. *The Canadian Journal of Sociology / Cahiers Canadiens de Sociologie*, 17(4), 468–469. <https://doi.org/10.2307/3341230>
- Longino, H. E. (1993). Subjects, power, and knowledge: Description and prescription in feminist philosophies of science. In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies*. Routledge.
- McGrath, A., & Jonker, A. (2024, October 8). What is AI interpretability?. IBM. <https://www.ibm.com/think/topics/interpretability>
- Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press. <https://doi.org/10.23943/princeton/9780691208411.001.0001>
- Rolin, K. (2006). The bias paradox in feminist standpoint epistemology. *Episteme*, 3(2-1), 125-136. <https://doi.org/10.3366/epi.2006.3.1-2.125>
- Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI & SOCIETY*, 39(4), 1585–1603. <https://doi.org/10.1007/s00146-022-01617-6>
- Sweet, P. L. (2020). Who knows? Reflexivity in feminist standpoint theory and Bourdieu. *Gender and Society*, 34(6), 922–950. <https://doi.org/10.1177/0891243220966600>
- Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology*, 35(4), Article 88. <https://doi.org/10.1007/s13347-022-00577-5>
- Wylie, A. (2004). Why standpoint matters. In S. Harding (Ed.), *The feminist standpoint theory reader: Intellectual and political controversies*. Routledge.
- Zowghi, D., & Bano, M. (2024). AI for all: Diversity and inclusion in AI. *AI Ethics*, 4(4), 873–876. <https://doi.org/10.1007/s43681-024-00485-8>