Philosophy & Technology (2022) 35:53
https://doi.org/10.1007/s13347-022-00548-w

RESEARCH ARTICLE

The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition

Rosalie Waelen[1] · **Michał Wieczorek**[2]

Received: 2 July 2021 / Accepted: 18 May 2022

Abstract
AI systems have often been found to contain gender biases. As a result of these gender biases, AI routinely fails to adequately recognize the needs, rights, and accomplishments of women. In this article, we use Axel Honneth's theory of recognition to argue that AI's gender biases are not only an ethical problem because they can lead to discrimination, but also because they resemble forms of misrecognition that can hurt women's self-development and self-worth. Furthermore, we argue that Honneth's theory of recognition offers a fruitful framework for improving our understanding of the psychological and normative implications of gender bias in modern technologies. Moreover, our Honnethian analysis of gender bias in AI shows that the goal of responsible AI requires us to address these issues not only through technical interventions, but also through a change in how we grant and deny recognition to each other.

Keywords Artificial intelligence · Gender bias · Ethics · Struggle for recognition · Axel Honneth

✉ Rosalie Waelen
r.a.waelen@utwente.nl

Michał Wieczorek
michal.wieczorek@dcu.ie

[1] Section of Philosophy, University of Twente, Enschede, Netherlands

[2] Institute of Ethics, Dublin City University, Dublin, Ireland

Published online: 03 June 2022

1 Introduction

In this paper, we use Axel Honneth's theory of recognition (Honneth, 1996) as a framework for the ethical analysis of gender biases in artificial intelligence (AI).[1] More specifically, we aim to show that biased AI systems can have detrimental effects on women's sense of self-worth. These effects on self-development are important ethical issues, but they are not addressed by existing AI ethics guidelines and principles (see Jobin et al., 2019, for an overview). Honneth's theory of recognition is valuable for helping us to understand the potential effects of AI on people's self-development. Moreover, Honneth's theory helps us to explain that bias is not only a technical problem, but also a social problem. Gender biases in AI, like racial bias and other forms of bias, indirectly stem from the social norms and practices that prevail in a society. Addressing the technical or material roots of AI's biases (for example by changing its training data or design) therefore does not suffice to ensure a move towards a future with more responsible AI. Responsible AI also requires us to reflect on society's structural relations of social recognition.

The structure of the paper is as follows. In Sect. 2, we briefly reconstruct Honneth's theory of recognition, which describes how an individual's development of self-confidence, self-respect, and self-esteem depends on others recognizing them with regard to love (e.g., emotional needs), rights (e.g., autonomy), and solidarity (e.g., societal contributions). Next, in Sect. 3, we apply Honneth's theory to analyze three examples of algorithmic and design biases in existing AI systems as instances of misrecognition of women. As a first example, we discuss how AI systems misrecognize women's voices and faces more often than men's, which makes it more difficult for women to use these technologies and enjoy their benefits (Sect. 3.1). Second, we show that AI can reinforce harmful stereotypes about women, which is a way of misrecognizing women's individuality, competency, and equality to men (Sect. 3.2). Third, we argue that AI and its developers too often ignore the diversity of women's values, perspectives, and needs, and thereby one-sidedly and reductively focus on the male point of view (Sect. 3.3). In Sect. 4, we conclude by discussing the

[1] Of course, Honneth's theory of recognition is not the only one as he himself was greatly influenced by G.W.F. Hegel's early writings on the subject. Moreover, other contemporary authors, most notably Judith Butler and Louis Althusser, developed their own understandings of the term which are much more ambiguous than what can be found in Honneth's work (cf. Lepold, 2021). In fact, Honneth has been criticized for presenting a much too optimistic view of recognition tied to the inevitability of social progress (Allen, 2021) and for failing to notice the negative dimension of recognition as leading to subjection and confining individuals into specific social roles established by the dominant groups (Butler, 2008; McNay, 2021). Furthermore, Honneth himself has moved on from the tripartite division of relations of recognition discussed in this article and developed an idea of antecedent recognition, understood as a precognitive awareness of the value of other people and of the meaning which they attributed to the physical world (Honneth, 2008). We do recognize the importance of the debates taking place among recognition theorists since Honneth's seminal *Struggle for Recognition* (Honneth, 1996) and hope that they will be reflected in future work on recognition and the ethics of technology. However, as the application of recognition theory to the analysis of ethical issues in AI is still a new approach within the philosophy of technology, we have decided to focus on Honneth's initial theory of recognition in the present paper. This exclusive focus allows us to introduce this approach and highlight its strength, without complicating our discussion with the highly nuanced and competing arguments put forward in the recognition debate since the initial publication of the book in question.

notion of "responsible AI" and offer some suggestions on how this ideal could be reinterpreted and realized in light of our Honnethian analysis of gender bias in AI.

A few more remarks are necessary before we can dive into the discussion. First of all, we understand AI as an umbrella term for a variety of techniques used to automate tasks—in line with today's usage of the term. What we will be discussing in particular are machine learning and deep learning algorithms, which are prone to inherit or develop biases. Secondly, it is important to note that Honneth's recognition theory deals with relations between individuals or social groups. The present work contributes to the debate on recognition and on the ethics of technology in that it explores a question that is significant to this day and age: What does recognition mean in a technological environment? Although this question has already been touched upon by a few authors, we believe that the topic of recognition deserves more attention in ethics of technology.[2]

Finally, by claiming that AI misrecognizes certain groups of users, we do not intend to attribute intentional agency or more autonomy than warranted to the technology. As will become clear in Sect. 3, most instances of AI's misrecognition of women stem from preexisting biases, such as prejudices held by developers or embedded in training datasets. In this sense, AI systems reproduce and amplify the relations of misrecognition already present in society. This idea is in line with recent work on mediated or mediational recognition in which a third party mediates the relation of recognition. Koskinen, (2019) provides a good example of this happening in property relations, when the recognition of the two parties is mediated by a reference to institutionalized law and the state upholding that law. In this situation, the two parties encounter and recognize each other under the influence of the rules codified in the legal system, but institutions such as courts can shape and embody relations of recognition as well. Similarly, in his debate with Nancy Fraser, Honneth argues that our experiences of (mis)recognition are mediated by language, cultural means of expression, and prevailing interpretations of individual circumstances. He also argues that the relations of recognition can be reflected in the material, economic sphere—in the division of labor and distribution of resources (Fraser & Honneth, 2003). Likewise, we believe that AI systems can both mediate our understanding of recognition and be an embodiment of existing societal structures of recognition. Moreover, from a phenomenological perspective, in their interactions with AI systems, users can have an impression of encountering AI as an autonomous entity and thus feel recognized or misrecognized by the system itself.[3] While we focus on analyzing AI-mediated misrecognition as an extension of (mis)recognition occurring in the society as a whole, we admit that the firsthand experience of users of AI systems can be different to our description. Hence, the discussion we present in this paper can apply to subjective feelings of misrecognition by AI and to relations of recognition mediated through AI.

[2] Namely: in a special issue of *Philosophy & Technology* on the topic of recognition (volume 33, issue 1); by Gertz, (2018), who uses Hegel's early work on the struggle for recognition (on which Honneth builds) to understand human–robot relations (Gertz, 2018); and by Waelen, (2022), who uses Honneth's theory of recognition in combination with Taylor's work on the topic in order to analyze the ethical implications of facial recognition technology.

[3] In terms of firsthand experience, users are likely to ignore the background infrastructure and design decisions behind the technologies they encounter on an everyday basis. This is particularly evident in our tendency to assign personality to our AI systems and AI-powered robots and treat them as independent entities. For example, blog and forum posts contain numerous instances of people naming their Roomba vacuum cleaners and treating them as house pets.

2 Honneth and the Struggle for Recognition

In his influential book *The Struggle for Recognition* (Honneth, 1996), Honneth outlines a social philosophy which focuses on recognition granted or refused to individuals and groups based on their needs, moral responsibility, and societal contributions. He highlights the intersubjective dimension of ethical life and provides an empirically justified account of how our relations with others influence our practical relation-to-self. In this section, we outline the three kinds of recognition and their normative implications, as discussed in *The Struggle for Recognition*. This theory then serves as a theoretical foundation for the next part of this paper, where we apply Honneth's framework of love, rights, and solidarity to three types of gender biases in AI systems.

According to recognition theory, our social relations shape our personality and identity, influencing the kinds of roles we adopt and the goals we pursue in everyday social practice. The recognition we receive or which we are denied while interacting with others is foundational for the development of individuals and the society as a whole. By interacting with others and by perceiving herself from their perspective, an individual develops a "practical relation-to-self" (Honneth, 1996, p. 92), which determines how she establishes her self-worth and sees her position in the society.

Love-based recognition refers to instances in which our physical and emotional needs are either affirmed or denied by others. Whereas Honneth discusses love primarily on the basis of the relation between a mother and her child, he also argues that it is "a pattern of interaction whose mature reappearance in adult life is an indication of successful affectional bonds to other people" (Honneth, 1996, p. 104). In a relation of love, one is recognized as an individual whose basic needs and feelings have value, and who consequently considers her own needs as valuable. According to Honneth, love is the primary mode of recognition and serves as the foundation for other social relations. Both during childhood and later in life, love makes it possible for an individual to develop "basic self-confidence" (Honneth, 1996, p. 129), which is necessary for human flourishing and a positive relation-to-self. While Honneth seems to associate violations in the sphere of love primarily with threats to physical integrity (e.g., physical abuse or rape), we believe that less extreme examples should also be considered important instances of misrecognition. In our view, love-related misrecognition could also be extended to situations in which an individual's feelings are routinely denied validity or not taken seriously, and when her unique needs are not considered by others. In this sense, we believe that misrecognition on the basis of love might most commonly manifest itself in the form of neglect rather than abuse.[4] As we demonstrate in the next section, AI systems routinely deny and exclude users' needs, perspectives, and values. The normative implication of such repeated violations of love-based recognition is that it can effectively undermine the development of *self-confidence* and, consequently, threaten individuals' general sense of self-worth.

Rights, the second relation of recognition discussed by Honneth, refers to an individual's autonomy and her capacity to make independent decisions that are recognized and respected by others. An individual is considered a bearer of rights only if she is seen

[4] We believe that this is in line with Honneth's discussion of parental love. A parent neglecting their child's need for emotional connection and nourishment might still be considered failing in their duties and threatening the child's self-confidence, even if such neglect is not a deliberate attempt at harm (in the way Honneth's examples of physical abuse and rape are).

as capable of entering into contracts and adhering to the social norms existing in a given community (with the added implication that she is also capable of shaping these norms). In this sense, recognition on the basis of rights is the recognition of an individual as a "morally responsible person" (Honneth, 1996, p. 118). Such a person is deserving of the respect of others, which in turn enables her to relate to herself with a sense of *self-respect*. The denial of recognition on the basis of rights can manifest itself in the political sphere (for example, through being denied group membership or even the right to vote), but for the purposes of this paper, we propose to consider rights primarily by focusing on the recognition individuals (do not) receive as persons capable of making their own decisions. Consequently, we believe it to be more appropriate to discuss (mis)recognition on the basis of rights by using the terms respect and disrespect. Although Honneth's notion of rights extends beyond the legal sphere and encompasses concepts such as agency, autonomy, or Kantian human dignity (see Kleinig & Evans, 2013), his emphasis on rights can lead readers to believe that this form of recognition deals exclusively with legal rights such as the right to vote or to own property. In our view, the term disrespect better illustrates the situations in which technological systems adopt a paternalistic tone, reinforce traditional gender roles, or deny women's equal worth to men, thus undermining women's capacity to make their own decisions and shape the trajectory of their lives (as well as impacting the societal recognition of these decisions).

Solidarity, the final relation of recognition described by Honneth, connects to individuals' societal contributions and their evaluation by other members of the society. This relation of recognition is often referred to as esteem. We commonly attribute more worth to certain professions and activities, which happens at the expense of others (compare, for example, the difference in the public perception of doctors and cleaners). In Honneth's recognition theory, the recognition one is given on the basis of one's abilities and individual traits influences her social standing and *self-esteem*. To recognize someone in terms of solidarity, it is important to consider her capable of engaging in socially valuable activities and believe she positively contributes to the society. In terms of AI, misrecognition on the basis of solidarity can manifest itself in biased recommendation systems that underappreciate the contributions of certain groups or reproduce gender stereotypes framing women's societal role as insignificant or subordinate to men's.

Recognition theory is not merely a descriptive tool, it allows for a normative societal analysis. By bringing instances of misrecognition to the surface, we are able to determine the often hidden and unquestioned moral beliefs existing within a society (Honneth, 2007) and reasons for social change (Honneth, 1996, pp. 131–139). In the next section, we demonstrate that an analysis of relations of recognition mediated by AI systems enables us to determine how AI contributes to gender injustices.

3 Misrecognition of Women by AI: Analyzing Some Examples

AI systems can be said to be biased in a variety of ways. According to Friedman and Nissenbaum, (1996), there are three categories of bias in computer systems: preexisting, technical, and emergent biases. Preexisting bias, first of all, entails that a system reproduces already existing human prejudices. These preexisting biases can

stem from the technical design of the system or from the training data on which the system was built. In the former sense, preexisting biases arise when the programming of a technology or its material aspects where designed in a way that reflects some prejudiced or non-inclusive beliefs held by developers (e.g., phones are designed to fit into the hand and the pocket of a man). In the latter sense, it is the case that non-inclusive input or training data result in biased algorithms that deliver low-quality, potentially discriminatory output. "Garbage in, garbage out," they say. Arguably, preexisting bias resulting from low-quality data is simultaneously a technical bias. Technical bias occurs when self-learning systems draw problematic conclusions from their training data and come to favor certain outcomes. Take the example of an algorithm that selects incoming university students. When trained on historical data, the system starts to associate being male or being White with being a good candidate. In this case, we can say the bias is a technical bias, as it drew wrong conclusions from the data, but it could also be framed as a preexisting bias to the extent that the training data was not representative of the population that the system would have to analyze and judge. Emergent biases, finally, arise when a system is used in a particular manner or context that was not intended or foreseen by the developers of the system.

Below we discuss three ways in which gender bias occurs in AI systems: the literal misrecognition of women's faces and voices, the reproduction of gender stereotypes, and the exclusion of female needs, perspectives, and values. They are mainly preexisting biases that stem from non-inclusive and non-representative training data, or from social prejudices held by tech-developers. We link each of the three instances of gender biases that we discuss to at least one of Honneth's relations of misrecognition. Although the three modes of misrecognition proposed by Honneth do not always correspond perfectly to the levels at which women are misrecognized by AI systems, we believe that Honneth's theory of recognition nevertheless offers a valuable framework to improve our understanding of the ethical and social impact of bias in AI systems.[5]

3.1 Misrecognizing Women, Literally

A first, most obvious example of AI's misrecognition of women is voice and face recognition systems that are less accurate in recognizing women than they are in recognizing men. Recognition, in this context, entails the identification of individuals or their traits through the sound of their voice, their appearance and facial features, or the things they say and their facial expressions. However, when recognition as identification fails, this can simultaneously be a case of misrecognition in the Honnethian sense.

[5] In fact, in Honneth's theory, all three modes of recognition are fundamental parts of individual and societal development and they are bound to influence each other. We believe that while the tripartite division is useful in analytical terms, real-world examples of misrecognition are likely to reflect an overall struggle for recognition encompassing all planes of societal relations. Our examples are a good illustration of this claim as they show recognitional imbalances occurring across all three modes of recognition.

Voice recognition, firstly, is a technology used in digital assistants like Siri, Alexa, and Google Assistant. It can imply the recognition of a particular individual on the basis of their voice, but it also refers to the mere recognition of what someone is saying (e.g., "Alexa, play some music"). The latter function is also referred to as "speech recognition". Studies have shown that speech recognition is less accurate when dealing with female voices (e.g., Tatman, 2017). This means that women are less likely to be understood or heard by a digital assistant and will more often face difficulties when interacting with these technologies. The same goes for people speaking with a distinct foreign or regional accent.

A second, similar example of literal misrecognition—i.e., misidentification—by AI is that of facial recognition systems. Like speech recognition, facial recognition has been found to perform worse on women than on men and, similarly, worse on dark-skinned individuals as opposed to light-skinned persons. Buolamwini and Gebru, (2018) evaluated three commercial facial analysis tools for the classification of gender and found that darker-skinned women were misclassified up to 34.7% of the time, while the error rate for lighter-skinned males was only 0.8%. While misrecognition of voices can be blamed on the fact that generally higher-pitched female voices are more difficult to process, the gap in recognition of female versus male faces (and light skin versus dark skin) is a result of a lack of inclusivity in training data sets. The former is an example of technical bias, while the latter reflects preexisting bias.

Women are literally being misrecognized by voice and facial recognition systems. As a result, they get to enjoy less benefits from these technologies and experience more difficulties when they interact with them. Simple ways of addressing this kind of misidentification, like lowering one's voice to sound more like men (Criado Perez, 2020), are flawed because they are simply impractical (who would remember or want to lower their voice every time they use their voice assistant?) and make women adapt to the male norm instead of changing that norm to become more inclusive. Furthermore, such solutions do not counter the deeper kind of misrecognition that women experience. That technologies are put on the market while they function significantly worse for female users—who are more often misidentified, misunderstood, or completely unnoticed by face and voice recognition systems—makes it seem like women are a less important or relevant audience to the tech industry. Voice and facial recognition systems first and foremost serve (White) men. Even if this was not intended by those who develop or implement the technology, and we assume that is the case, the rollout of technologies that are much more difficult to use by women makes it appear as if women are a less important group of users. This is a misrecognition of women's needs, that is, misrecognition on the basis of love, because their needs as users are not met. Moreover, misrecognition by voice and face recognition systems is also an instance of misrecognition on the basis of solidarity. Women are treated as second-rate users of the technology, and they are routinely misclassified and misunderstood by technologies designed to assist and empower them, which is reflective of the (lack of) acknowledgement for women's societal importance. From understanding the misidentification of women's voices and faces as misrecognition in the Honnethian sense follows that misidentification by voice and facial recognition systems threatens women's development of self-confidence and self-esteem.

Of course, it is a strong claim that the misidentification of women can hamper their ability to develop self-confidence and self-esteem. However, Honneth's theory of recognition gives us reason to argue that this might be the case and, moreover, motivates future research to empirically investigate to what extent women's recurring experience that they are a less important group of users than men indeed affects their ability to develop self-esteem, self-confidence, and an overall sense of self-worth. Existing research already suggests that technology can have such an effect on women's self-development. Several studies have already been conducted to show the effect of social media on people's self-image. For instance, Jiang and Ngien, (2020) investigated the effect that social media platform Instagram had on the social anxiety and self-esteem of its users. They found that the social comparison that Instagram encourages increased user's social anxiety and significantly decreased self-esteem. Studies have also been conducted regarding the effect of fitness-tracking devices and apps on people's self-image and they have demonstrated that users failing to achieve the levels of activity recommended by the algorithms report feelings of anxiety, powerlessness, and lowered self-esteem (Kristensen et al., 2021; Lupton, 2013; Owens & Cribb, 2019). Hence, it seems true and clear that AI can have negative implications for individuals' self-confidence and self-esteem, making Honneth's framework a relevant and valuable tool for identifying and analyzing these effects.

3.2 Reinforcing Stereotypes About Women's Traits and Role in Society

Another way in which AI systems misrecognize women is by reinforcing problematic, offensive, or simply inaccurate stereotypes about them. Sadly, there is a big pool of examples of gender stereotypes in AI applications as well as in other types of technologies. A first example is the fact that automated credit calculation systems can assign women a lower credit score and credit card limits than their male counterparts, suggesting that women cannot handle money well (Vigdor, 2019). Furthermore, AI systems used in human resources are found to ignore women in hiring and promotion decisions (Dastin, 2018). Women are also shown different job advertisements, thereby missing out on positions stereotypically perceived as male and traditionally better paid, which again contributes to the entrenchment of the unequal position of women in the professional world (Imana et al., 2021; Wachter-Boettcher, 2017). Algorithms used by social networking sites routinely (mis)attribute users' gender through inferences made on the basis of sexist stereotypes and binary gender roles, while not acknowledging that users express their gender and sexuality in diverse ways (Fosch-Villaronga et al., 2021). Fertility tracking apps have been perceived as patronizing because of their highly gendered design and inherent assumption that women using the app would want to get pregnant (Hall, 2017; Kressbach, 2019). Such apps reinforce the stereotypical expectation that a woman should have, and should want to have, children. Similarly, other apps and devices related to sexuality or intimacy reinforce existing gender stereotypes and endorse roles traditionally occupied by men and women in romantic relationships. For example, these apps and devices equate male sexual performance with physical exertion or associate male reproductive and romantic success with the number of female sexual partners

one has had, which means that apps can facilitate the objectification of women (cf. Danaher et al., 2018; Lupton, 2015). Finally, the aforementioned digital voice assistants usually come with a female voice and name. As users anthropomorphize these assistants, they learn to relate the female gender to the servile attitude and role of digital assistants (Specia, 2019) and potentially internalize the traditional image of women as subservient to others and attending to their needs.

These stereotypes first of all misrecognize women's individuality—not all women like the color pink, wish to bear children, or pursue traditionally feminine careers such as that of a kindergarten teacher or nurse. Women are unique individuals who cannot be reduced to a limited set of traits. Secondly, by relying on stereotypes, AI systems fail to recognize women as men's equals, especially as they promote a view of women as subservient to men. This misrecognition of women through gender stereotypes in AI relates to all three types of misrecognition described by Honneth. Stereotypes fuel the misrecognition of women on the basis of love, by perceiving women as merely being there to serve others rather than individuals with their own desires and needs. AI systems disrespect women, because the stereotypes embodied by AI can have a constitutive effect on women's identities and can effectively reduce the diversity of roles, behaviors, and life choices that are pursued by women. By being constantly confronted with gender stereotypes through AI, women could be encouraged or even pressured to function and develop (consciously or unconsciously) in conformity with these stereotypes. Digital assistants, for instance, teach women that they need to be kind, servile, and flirtatious, no matter how they are spoken to, which can keep women from developing the kind of self-respect that moves them to speak up or defend themselves when needed.

Finally, some of the mentioned stereotypes we see reflected in AI resemble a lack of esteem. For example, AI systems used to assist hiring and promoting decisions do not esteem women's societal contributions adequately when they do not value time taken off for bearing and raising children, as it would not only keep women from having a successful career, but also neglect the societal value of these naturally and traditionally feminine roles.

Taken together, these examples of misrecognition through gender stereotypes can have some considerable negative effect on women's development of self-worth, as well as their actual possibility to employ their skills and pursue their goals. Admittedly, these stereotypes exist outside of the AI-mediated sphere as well. The problem is that their negative implications are kept in place by AI, often without our awareness. In a way, this makes AI not exactly a *modern* technology. Moreover, the growing prevalence of AI systems in all walks of life, as well as the opacity and the often perceived objectivity of the decisions made by AI systems, might exacerbate the frequency and significance of the negative impacts of gender stereotypes.

3.3 Excluding Female Needs, Perspectives, and Values

Young, white, and often affluent men make up the majority of the workforce in many technology companies today (Richter, 2021). This gives them a much greater influence on the development of technologies than other groups have. Consequently, the

needs, perspectives, and values of those other groups, among which are women, at times end up being ignored in the design of AI systems and thus not sufficiently represented in the technologies that emerge in our society. Moreover, as the male point of view is overrepresented in the development of some technologies, that view sets the standard and women are forced to adjust to products that are tailored towards men. This male point of view can include unwanted stereotypes about women, as discussed in Sect. 3.2. However, the bias resulting from a lack of inclusion of women in AI does not necessarily stem from the prejudices that technology developers hold about women and can therefore be classified as preexisting as well as emergent bias. Biases can arise despite good intentions of developers, when AI replicates harmful principles or is deployed in a context that makes it more likely to arrive at biased outcomes (Friedman & Nissenbaum, 1996).

Development teams with inadequate female representation are likely to overlook features that could be relevant to women. This was for example the case with Apple's Health toolkit, which for a long time did not allow users to track menstruation (Duhaime-Ross, 2014). Moreover, due to their one-sided perspective, developers can also fail to anticipate the plausible, different ways in which female users utilize their AI-powered products. This, in turn, might influence the accuracy of the device and the quality of the provided services. For instance, a woman using a fitness-tracking application might receive less accurate recommendations and predictions as the app has been designed with the assumption that smartphone devices, used as a data source, will be carried in a pocket of the user's clothes instead of in a handbag (Criado Perez, 2020). Moreover, apps related to sexuality often replicate the male view of sexual relationships by framing sex in connection with male-defined and male-centered parameters such as intensity and physical exertion (for example, monitoring the intensity of thrusts and the number of calories burned by a male partner), or by excessively supplying women with information and recommendations dealing with medical issues and risk, rather than, for example, satisfaction (Danaher et al., 2018; Lupton, 2015).

A similar example of gender bias in AI and other digital technologies is the often heard complaint that female profiles on social media are less likely to be recommended to other users, which gives men an unfair advantage in likes, followers, or even income they receive as a consequence of their social media posts (Beard et al., 2020). As Cobbe, (2020) writes: social media platforms are "marginalising women and LGBT people by removing or restricting their communications". Hence, women have to deal with a glass ceiling even when choosing to make a living as influencers on social media, and their ability to reach an audience potentially interested in their perspective is limited in comparison to men.

These examples of a lack of inclusion of women's needs, perspectives, and values again represent all three kinds of recognition highlighted by Honneth. The recognition of an individual's needs, perspective, and values as important and their reflection in the design of technologies produced on a massive scale provides said individual with affirmation that she herself and her wants and desires matter for other members of the society and that their fulfillment is worth the required effort. This can be interpreted as recognition on the basis of love. Without this kind of acknowledgment, an individual's ability to develop a sense of self-confidence is severely

limited and can lead to her downplaying or ignoring her individual needs as not worthy of anyone's attention.

Moreover, the glass ceiling discussed in the context of social media algorithms points to disrespect.[6] By limiting women's discoverability and reach on social media platforms, AI infringes on women's opportunities (for example their chance to make a living as influencers) and limits the variety of choices effectively available to them, thus reducing their autonomy. Moreover, AI and its functionalities can be seen as reflecting which kinds of choices and life paths are socially recognized as viable and lying within the scope of (female) users' decision-making capabilities. If a society does not believe women should have agency regarding their fertility, it stands to reason that it would not create possibilities for women to take control over their reproductive health. In this sense, lack of features that could be reasonably expected by some groups might be interpreted not merely as a careless omission, but as a sign of a deeper, structural disrespect of women.

Similar to the example discussed in Sect. 3.1, the failure to reflect women's needs in technology products suggests that women are not users worth designing for and that their needs, perspectives, and values do not warrant the inclusion in the design process of AI systems. As it seems, the female user group is not perceived as important enough to justify additional effort on the part of tech companies. This can make women believe that they are not recognized as relevant on the societal level, which can significantly damage their self-esteem. The existing practice of ignoring the many and diverse views and values of women in technology design is thus a case of misrecognition on the basis of solidarity.[7]

4 Responsible AI in Light of Women's Struggle for Recognition

The misrecognition of women's needs, accomplishments, and rights (that is, the disrespect of women) is a longstanding societal injustice that, as we have shown in the previous section, is now exacerbated by AI systems. We maintain that Honneth's theory of recognition offers a valuable, new understanding of the normative implications of biased AI systems. Analyzing gender biases in terms of misrecognition reveals their potential negative impact on women's self-development and self-worth. What is left is to discuss how to tackle the misrecognition of women by AI in order to achieve responsible AI. As we will argue below, Honneth's philosophy once again proves to be a valuable tool for the ethics of AI.

[6] The term glass ceiling could also refer to a gendered pay gap, which would imply a lack of esteem (i.e., the undervaluation, in monetary terms, of women's contributions to the society). However, we decided to focus on respect because the misrecognition discussed in this paragraph infringes women's freedom to pursue a specific career rather than impacting their compensation in that career.

[7] We are aware that this issue hints at a tension between reliance on stereotypes and an actual recognition of the wide variety of views and values endorsed by women, which might not be possible from a technical standpoint as AI systems need to depend on wide and potentially reductive categories. Nevertheless, we wanted to highlight the necessity of making attempts at considering and designing for the diversity of potential users of new technologies.

4.1 How to Counter Misrecognition by AI?

As each of the discussed instances of gender bias has different causes, they need to be tackled in different ways. We offer some suggestions for each of the three instances of bias and the misrecognition they give rise to. The first type of misrecognition we discussed in Sect. 3, the misidentification of women's faces or voices, is mainly caused by insufficient inclusion of women's voices and faces in training data (Buolamwini & Gebru, 2018). Therefore, this form of misrecognition could be countered by the use of more appropriate data sets in the creation of facial recognition or speech recognition models. However, the misrecognition of women's voices may also have to do with the technical challenge of recognizing higher-pitched voices. In that case, more research needs to be invested into developing technology that can adequately capture women's voices.

The second type of misrecognition, namely gender stereotypes in AI, can be countered by making products less gendered and more representative of the diversity of ways in which users express their gender and sexuality. For example, digital assistants or robots could come without a gendered name and default voice. Users would then have to name the device and choose the voice themselves at the start. Another way to prevent stereotypes is to give users more opportunities to give input about their profiles and preferences and thereby exercise more control over personalized services.

The third type of misrecognition we discussed results from reliance on reductive generalizations about femininity and the omission of the diversity of women's needs, values, and perspectives. One way to avoid this is simply by including more women in the development of AI systems and ensuring that they have an equal say in the process. As a consequence of such inclusion, female needs, perspectives, and values will influence the decisions made by design teams.

With these and similar measures, we could expect gender biases in AI, and thus misrecognition of women by AI systems, to decrease significantly. However, these measures merely address the technical or design-related causes of gender biases in AI, such as the non-inclusive data sets or development teams that lead to biased AI systems. They do not transform the social structures that enabled the creation of biased data sets or led to the non-inclusive composition of the development teams in the first place. In the next sub-section (Sect. 4.2), we argue that we need to address the social roots of AI's biases towards women, if we really want to solve the problem and realize a future with responsible AI.

4.2 Realizing Responsible AI

Based on our Honnethian analysis of gender biases in AI systems, we offer three reflections on how to achieve responsible AI, which go beyond addressing the immediate technical or design-related causes of bias. Before doing so, it should be noted that there are, broadly, two different ways to interpret the much-heard call for "responsible AI". First of all, there is the question regarding the (legal) liability or accountability of intelligent agents (e.g., Santoni de Sio & Mecacci, 2021). Existing

AI ethics guidelines discuss different actors, such as technology developers, designers, companies, and institutions as "being responsible and accountable for AI's actions and decisions" (Jobin et al., 2019, 7). Secondly, discussions about responsible AI can also refer to how AI is developed and put to use.[8] Virginia Dignum describes this understanding of responsible AI as follows:

> (…) in the same way as we have the choice to use organic apples to make our pie, in AI we also have the choice to use data that respects and ensures fairness, privacy, transparency, and all the other values that we hold dear. This is what responsible AI is about – the decisions taken concerning the scope, the rules and the resources that are used to develop, deploy, and use AI systems (Dignum, 2020, 217).

In other words, this second interpretation of the term "responsible AI" refers to the processes that will lead to AI that complies with ethical principles. It is this interpretation of responsible AI that we are concerned with here.

Our first reflection on achieving responsible AI is that this goal cannot entail a continuation of the existing, flawed practices that made the development and adoption of biased technology possible in the first place. Our Honnethian analysis of AI's gender biases showed that these biases are symptoms of women's ongoing struggle for recognition in society. A similar argument can be made for racial equality: Racial biases in AI reflect the ongoing struggle of non-Whites in dominantly White societies. The prevalence of technology that works significantly worse for non-males and non-Whites does not point to isolated incidents caused by inadequate datasets or inattentive developers but shows that our social practices fail to live up to our proclaimed normative commitments. Although certainly important, technical interventions and practical solutions alone cannot bring us closer to the goal of responsible AI by themselves. We need to change our social norms and practices to take more seriously how we grant and deny recognition to each other. In our view, a critique and revision of current social norms and their embodiment in actual practices is just as important for the goal of responsible AI as more focused, localized interventions into specific data sets, algorithms, or design processes.[9]

Secondly, our Honnethian analysis shows that the ethical analysis of gender bias in AI can benefit from the consideration of the social, political, and historical context of AI's biases. As already pointed out, AI's biases are not isolated incidents. A thorough understanding of the social, historical, and political roots of these biases is needed if

[8] The term "responsible AI" is for example discussed by Google: https://ai.google/responsibilities/responsible-ai-practices/ (accessed January 17th, 2022).

[9] Even though Honneth emphasizes the need of immanent critique of the society (that is, on the basis of values already manifested within it), he still requires us to imagine the direction in which we are headed and refer in our analysis of social struggles to a hypothetical endpoint of relations of recognition (see Honneth, 1996, 171–179). We believe that this hypothetical, future-oriented outlook requires us to not only improve existing social practices and their associated norms, but also imagine new ones (which is arguably reflected in Honneth's discussion of democracy in which he draws on Dewey to argue that a "far-reaching, radical redefinition" of our social systems and values might be necessary if we want to live up to our normative commitments (Honneth, 1998, 780)).

we want to provide a complete understanding of the ethical issues caused by biased AI systems. Popular approaches in AI ethics, such as AI ethics principles and guidelines, that ought to realize responsible AI, demand little consideration for structural power dynamics that allow these issues to occur. Our analysis of women's struggle for recognition in the context of AI has shown that considering the social, political, or historical context in which AI is developed and used can help us identify and understand some specific ethical implications that the technology might have. Therefore, we conclude that an analysis of gender bias in AI should not merely be an ethical analysis, but ought to be complemented by a social analysis—namely an analysis of the social norms or the power structures in a society that cause certain ethical issues to occur in a systematic manner. Moreover, the focus on recognition as the guiding concept of our analysis allows us to discuss not only who is misrecognized, but also which groups stand behind this misrecognition (e.g., designers of AI or the people who created the data set responsible for the bias embedded in the system). In this sense, the analysis of relations of recognition makes it possible to uncover the power structures within which technologies function, and this is something that has so far been largely ignored by the AI ethics debate (cf. Crawford, 2021).

Third, current guidelines to realize responsible or ethical AI are focused on the harms and benefits brought by AI. Honneth's theory of recognition teaches us that the implications of misrecognition by AI cannot always be felt or observed directly. Our analysis of gender bias has shown that AI not only has the power to influence our behavior directly, but also constitutes who we become and how we are able to express ourselves over time. Misrecognition, especially when it occurs structurally, has a negative impact on people's development of self-confidence, self-respect, and self-esteem. In other words, even though misrecognition might not appear to have significant harms in the moment that it occurs, its constitutive effect on a person's self-worth can be harmful in the long run. Therefore, we conclude that truly responsible AI implies considering not only direct, observable harms and benefits, but also the long-term effects of AI on people's self-development.

Moreover, AI's power to shape people's identity does not merely involve negative effects. It is interesting to explore how AI's constitutive power can be put to use to positively influence people's self-development. One potential example of this is Netflix's personalization algorithms (Plummer, 2017). Netflix personalizes user profiles in two ways: by highlighting and recommending the movies and series that are inferred to be most in line with a user's interests and preferences, and by displaying a scene or image of the movie or series that best fits the user's profile. For example, if a user is expected to like content with a strong female lead, Netflix not only suggests such content, but also tweaks the thumbnails of the films and series it suggests to highlight female characters. This approach might of course result in flawed suggestions based on harmful stereotypes (Zarum, 2018), but when applied successfully, Netflix's approach can highlight the accomplishments of often marginalized groups and account for their diverse needs and desires. Personalization of content and thumbnails can help to send female users the message that women can also play important, interesting parts and that movies or series about women, or about the user's interests, are valuable. Hence, Netflix's algorithms can positively shape women's sense of self-worth by recognizing their interests, needs, and value to society.

However, highlighting female characters is not enough if leads in movies and series continue to be disproportionately played by men. Moreover, although it can positively affect people's sense of self-worth, Netflix's nudging of users to consume specific content can also be seen as an infringement of users' autonomy. Netflix's personalization algorithms are ambiguous as they can co-determine the kinds of identity and expressions of identity available to users and influence their self-development also in a negative manner. But rather than advocating for an outright rejection of such systems because of the potential dangers associated with AI's constitutive power, we argue that it would be more valuable to explore how this constitutive power can be used to our advantage. By analyzing how AI technologies impact our relations, possibilities, and vocabularies of recognition, we can attempt to harness AI's power to shape people's views in ways that would strengthen, rather than inhibit, people's development of self-worth.

Author Contribution Both authors contributed equally to the creation of this article.

Funding Both authors received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 813497.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

Allen, A. (2021). Recognizing ambivalence: Honneth, Butler, and philosophical anthropology. In H. Ikäheimo, K. Lepold, & T. Stahl (Eds.), *Recognition and ambivalence* (pp. 99–127). Columbia University Press.

Beard, L., Dunn, J., Huang, J. &, Krivkovich, A. (2020). *Shattering the glass screen.* McKinsey & Company, Technology, Media and Telecommunications. Retrieved from https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/shattering-the-glass-screen. Accessed 1 June 2022.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability, and Transparency*, 77–91.
Butler, J. (2008). Taking another's view: Ambivalent implications. In M. Jay (Ed.), *Reification: A new look at an old idea* (pp. 97–119). Oxford University Press.
Butler, J. (2021). Recognition and the social bond: A response to Axel Honneth. In H. Ikäheimo, K. Lepold, & T. Stahl (Eds.), *Recognition and ambivalence* (pp. 31–53). Columbia University Press.
Cave, S., & Dihal, K. (2020). The whiteness of AI. *Philosophy and Technology, 33*(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6
Cobbe, J. (2020). Algorithmic censorship by social platforms: Power and resistance. *Philosophy and Technology*. https://doi.org/10.1007/s13347-020-00429-0
Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
Criado Perez, C. (2020). *Invisible women*. Vintage.
Danaher, J., Nyholm, S., & Earp, B. D. (2018). The quantified relationship. *The American Journal of Bioethics, 18*(2), 3–19. https://doi.org/10.1080/15265161.2017.1409823
Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUKKCN1MK08G. Accessed 1 June 2022.
Dignum, V. (2020). Responsibility and artificial intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 215–231). Oxford University Press.
Duhaime-Ross, A. (2014). Apple promised an expansive health app, so why can't I track menstruation? *The Verge*. Retrieved from https://www.theverge.com/2014/9/25/6844021/apple-promised-an-expansive-health-app-so-why-cant-i-track. Accessed 1 June 2022.
Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., & Custers, B. H. M. (2021). A little bird told me your gender: Gender inferences in social media. *Information Processing and Management, 58*(3), 102541. https://doi.org/10.1016/j.ipm.2021.102541
Fraser, N., & Honneth, A. (2003). *Redistribution or recognition?* Verso.
Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *Computer Ethics, 14*(3), 215–232. https://doi.org/10.4324/9781315259697-23
Gertz, N. (2018). Hegel, the struggle for recognition, and robots. *Techné Research in Philosophy and Technology, 22*(2), 138–157.
Google (Date Unknown). *Responsible AI practices*. Retrieved from https://ai.google/responsibilities/responsible-ai-practices/. Accessed 17 Jan 2022.
Hall, M. (2017). The strange sexism of period apps. *Vice*. Retrieved from https://www.vice.com/en_us/article/qvp5yd/the-strange-sexism-of-period-apps. Accessed 1 June 2022.
Honneth, A. (1998). Democracy as reflexive cooperation: John Dewey and the theory of democracy today. *Political Theory, 26*(6), 763–783.
Honneth, A. (2007). *Moral consciousness and class domination*. Polity Press.
Honneth, A. (1996). *The struggle for recognition: The moral grammar of social conflicts*. MIT Press.
Honneth, A. (2008). *Reification: A new look at an old idea* (M. Jay, Ed.). Oxford University Press.
Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. *Proceedings of the Web Conference, 2021*, 3767–3778. https://doi.org/10.1145/3442381.3450077
Jiang, S. & Ngien, A. (2020). The effects of Instagram use, social comparison, and self-esteem on social anxiety: A survey study in Singapore. *Social Media + Society*. 1–10. https://doi.org/10.1177/2056305120912488
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-019-0088-2
Kleinig, J., & Evans, N. G. (2013). Human Flourishing, Human Dignity, and Human Rights. *Law and Philosophy, 32*(5), 539–564. https://doi.org/10.1007/s10982-012-9153-2
Koskinen, H. J. (2019). Mediated recognition: Suggestions towards an articulation. In M. Kahlos, H. J. Koskinen, & R. Palmén (Eds.), *Recognition and religion: Contemporary and historical perspectives* (pp. 34–50). Routledge, Taylor & Francis Group.
Kressbach, M. (2019). Period hacks: Menstruating in the big data paradigm. *Television and New Media, 00*, 1–21.
Kristensen, D. B., Kuruoglu, A. P., & Banke, S. (2021). Tracking towards care: Relational affordances of self-tracking in gym culture. *Sociology of Health and Illness, 43*(7), 1598–1613. https://doi.org/10.1111/1467-9566.13352

Lepold, K. How should we understand the ambivalence of recognition? Revisiting the link between recognition and subjection in the works of Althusser and Butler. In H. Ikäheimo, K. Lepold, & T. Stahl (Eds.), *Recognition and ambivalence* (pp. 129–59). Columbia University Press.

Lupton, D. (2013). Quantifying the body: Monitoring and measuring health in the age of mHealth technologies. *Critical Public Health, 23*(4), 393–403. https://doi.org/10.1080/09581596.2013.794931

Lupton, D. (2015). Quantified sex: A critical analysis of sexual and reproductive self-tracking using apps. *Culture Health and Sexuality, 17*(4), 440–453.

McNay, L. (2021). Historicizing recognition: From ontology to teleology. In H. Ikäheimo, K. Lepold, & T. Stahl (Eds.), *Recognition and ambivalence* (pp. 69–97). Columbia University Press.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Owens, J., & Cribb, A. (2019). 'My Fitbit thinks i can do better!' Do health promoting wearable technologies support personal autonomy? Philosophy and Technology, 32(1), 23–38. *Scopus*. https://doi.org/10.1007/s13347-017-0266-2

Plummer, L. (2017). This is how Netflix's top-secret recommendation system works. *Wired*. Retrieved from https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like. Accessed 1 June 2022.

Richter, F. (2021). Women's representation in big tech. *Statista*. Retrieved from https://www.statista.com/chart/4467/female-employees-at-tech-companies/. Accessed 18 Jan 2022.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology*. https://doi.org/10.1007/s13347-021-00450-x

Specia, M. (2019). Siri and Alexa reinforce gender bias, U.N. Finds. *The New York Times*. Retrieved from https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html. Accessed 1 June 2022.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. Proceedings of the First Workshop on Ethics in Natural Language Processing (April 4th 2017 Valencia, Spain), pp. 53–59.

Verbeek, P. (2011). *Moralizing technology: Understanding and designing the morality of things*. The University of Chicago Press.

Vigdor, N. (2019). Apple card investigated after gender discrimination complaints. *The New York Times*. Retrieved from https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. Accessed 1 June 2022

Wachter-Boettcher, S. (2017). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech* (1st ed.). W.W. Norton & Company.

Waelen, R. (2022). The struggle for recognition in the age of facial recognition technology. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00146-8

Zarum, L. (2018). Some viewers think Netflix is targeting them by race. Here's what to know. *The New York Times*. Retrieved from https://www.nytimes.com/2018/10/23/arts/television/netflix-race-targeting-personalization.html. Accessed 1 June 2022.