cogitatio

SOCIAL INCLUSION

Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination

Edited by Derya Ozkul

Volume 12

2024

Open Access Journal

ISSN: 2183-2803

Social Inclusion, 2024, Volume 12
Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination

Published by Cogitatio Press
Rua Fialho de Almeida 14, 2º Esq.,
1070–129 Lisbon
Portugal

Design by Typografia®
http://www.typografia.pt/en/

Cover image: © IR_Stone from iStock

Academic Editor
Derya Ozkul (University of Warwick)

Available online at: www.cogitatiopress.com/socialinclusion

Table of Contents

Social Inclusion
2024 • Volume 12 • Editorial 8942
https://doi.org/10.17645/si.v12.8942

cogitatio

EDITORIAL

Open Access Journal

Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination

Derya Ozkul

Department of Sociology, University of Warwick, UK

Correspondence: Derya Ozkul (derya.ozkul@warwick.ac.uk)

Submitted: 15 July 2024 **Published:** 30 July 2024

Issue: This editorial is part of the issue "Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination" edited by Derya Ozkul (University of Warwick), fully open access at https://doi.org/10.17645/si.i236

Abstract

This thematic issue explores the applications of artificial intelligence-based technologies and their potential for producing discriminatory and biased outcomes based on ethnicity, religion, and gender. This thematic issue adds to the ongoing debate with theoretical and empirical studies and a commentary that examine the topic from various perspectives. This editorial discusses the key themes highlighted in the studies and presents the findings of the different contributions to this collection.

Keywords

algorithms; artificial intelligence; automated decision-making systems; bias; discrimination

Numerous studies have shown that artificial intelligence (AI) technologies can produce biased outcomes due to their design and the existing inequalities in society (Benjamin, 2019; Broussard, 2023; Joyce et al., 2021). For example, hiring algorithms have been found to discriminate against women and individuals with minority names. Gender and racial discrimination have also been identified in internet search platforms (Noble, 2018), targeted ads and social media posts, credit scoring, insurance provision (O'Neil, 2016), and facial recognition technologies, especially in their accuracy with black female faces (Buolamwini & Gebru, 2018). Biased credit scoring and insurance provisions could worsen inequality, while biased algorithms in social media for voter targeting could threaten democracy (O'Neil, 2016). The lack of transparency makes it difficult for outsiders to understand how the systems were designed and by whom, highlighting the need for "data feminism" to analyse inherent inequalities leading to biases (D'Ignazio & Klein, 2020). Additionally, the lack of transparency in how these automated systems function creates challenges for users in identifying and contesting biased outcomes.

This thematic issue builds another brick into the debate, with studies exploring it from different angles. In the first article, author Carsten Orwat analyses the risk of discrimination in relation to broader threats to

fundamental rights and points out the shortcomings of current anti-discrimination laws in addressing algorithmic discrimination (Orwat, 2024). The use of AI and automated decision-making systems (ADMs), he argues, present challenges for data subjects who may find it difficult to recognise unfair treatment and gather enough evidence to challenge discriminatory practices caused by algorithms. As a result, Orwat contends, data subjects may face significant obstacles in taking legal action. He further explores how discrimination is linked to potential violations of human dignity.

Orwat conceptualises human dignity not as monolithic, but as multifaceted. Looking at the decisions of the German Federal Constitutional Court and developments of constitutional law in Germany, he shows that the concept of human dignity is already concretised in law, suggesting that the human being cannot be treated like an object and has the right to equal treatment and the free development of personality. As his analysis of past court rulings shows, the Federal Constitutional Court has concretised its interpretation of the right to human dignity and personality in relation to the risks posed by technologies. He then proceeds with analysing violations of human dignity concerning specific factors, namely severe and structural discrimination that has an impact on groups of individuals that are not regarded and treated with equal moral worth; lack of understanding around differentiating criteria from data subjects' perspectives; data subjects being treated as objects instead of individuals with respect and as whole persons; externalisation of identity determination and thereby elimination of self-determination, as well as the absence of meaningful informed consent in these automated processes. His article emphasises the importance of treating individuals with respect in complex human and machine interactions.

Larsson et al. (2024) present a related yet empirical case study in their article, exploring discrimination in the employers' use of AI and ADMs. While previous research has extensively examined potential biases arising from the use of AI in recruitment, this study shifts its focus to the perceptions of recruiters regarding AI and ADMs. Larsson and colleagues conducted an analysis of how recruiters comprehend AI and ADMs based on data they collected from prominent recruitment agencies and employers in Sweden. Their findings indicate a general lack of awareness among recruiters regarding the everyday integration of AI and ADM in their workplace. This lack of awareness may stem from a failure to recognise numerous common technological applications, such as searching for information on Google and encountering advertisements, in connection to algorithms and automated recommendation systems. It may also be possible that participants in the study may perhaps perceive AI as a relatively novel concept. However, the study's results demonstrate that as AI and ADMs progressively pervade everyday life, they become normalised and even disregarded by their users. This phenomenon is notably pronounced in cases where the use of these technologies is indirect and not perceived as central to the recruiters' responsibilities despite being extensively used.

The article by Inga Ulnicane explores another specific case study, this time focusing on how debates around AI frame emerging problems from the use of AI (Ulnicane, 2024). The author focuses on analysing the findings of four specific reports that she considers to be high-profile by using a framing approach and highlights their consistent emphasis on the dearth of diversity in the AI workforce, instances of AI-induced discrimination, and the inadequacies of nascent AI policies and guidelines. In addition to identifying common themes in these reports, Ulnicane also underscores their contextual specificity, raising questions about who has the power to identify problems and formulate recommendations within this domain. Indeed, in line with prevailing practices, leading reports in any given field often exhibit ties to the geopolitical powers underwriting their production.

Finally, in a short commentary, Pramod K. Nayar discusses whether or not we can think of machines as entities in their own right (Nayar, 2024). He prompts us to question who possesses the authority to withhold the entitlements of machines, especially in a world where humans are increasingly integrated with technology. Specifically, Nayar raises the question of whether robots, capable of performing tasks traditionally associated with humans and demonstrating cognitive and emotional capabilities, including biases, should be excluded from the institution of citizenship. He contends that withholding rights and responsibilities from robots could in itself be a form of discrimination. This provocative piece urges us to consider whether humans are the sole arbiters of the world's governing principles or if the paradigm of human-robot interactions should be viewed as intertwined rather than separate when considering these principles.

Conflict of Interests

The author declares no conflict of interests.

References

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency in Proceedings of Machine Learning Research*, *81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius*, *7*. https://doi.org/10.1177/2378023121999581

Larsson, S., Merricks White, J., & Ingram Bogusz, C. (2024). The artificial recruiter: Risks of discrimination in employers' use of ai and automated decision-making. *Social Inclusion*, *12*, Article 7471. https://doi.org/10.17645/si.v12.7471

Nayar, P. K. (2024). How to include artificial bodies as citizens. *Social Inclusion*, *12*, Article 8337. https://doi.org/10.17645/si.v12.8337

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

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

Orwat, C. (2024). Algorithmic discrimination from the perspective of human dignity. *Social Inclusion*, *12*, Article 7160. https://doi.org/10.17645/si.v12.7160

Ulnicane, I. (2024). Intersectionality in artificial intelligence: Framing concerns and recommendations for action. *Social Inclusion*, *12*, Article 7543. https://doi.org/10.17645/si.v12.7543

About the Author

Derya Ozkul (PhD) is an assistant professor in sociology at the University of Warwick and a research associate at the University of Oxford's Refugee Studies Centre.

Social Inclusion
2024 • Volume 12 • Article 7160
https://doi.org/10.17645/si.v12.7160

cogitatio

ARTICLE

Open Access Journal

Algorithmic Discrimination From the Perspective of Human Dignity

Carsten Orwat

Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Germany

Correspondence: Carsten Orwat (orwat@kit.edu)

Submitted: 8 May 2023 **Accepted:** 15 February 2024 **Published:** 13 May 2024

Issue: This article is part of the issue "Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination" edited by Derya Ozkul (University of Warwick), fully open access at https://doi.org/10.17645/si.i236

Abstract

Applications of artificial intelligence, algorithmic differentiation, and automated decision-making systems aim to improve the efficiency of decision-making for differentiating persons. However, they may also pose new risks to fundamental rights, including the risk of discrimination and potential violations of human dignity. Anti-discrimination law is not only based on the principles of justice and equal treatment but also aims to ensure the free development of one's personality and the protection of human dignity. This article examines developments in AI and algorithmic differentiation from the perspective of human dignity. Problems addressed include the expansion of the reach of algorithmic decisions, the potential for serious, systematic, or structural discrimination, the phenomenon of statistical discrimination and the treatment of persons not as individuals, deficits in the regulation of automated decisions and informed consent, the creation and use of comprehensive and personality-constituting personal and group profiles, and the increase in structural dominance.

Keywords

algorithmic discrimination; artificial intelligence; automated decision-making; development of personality; generalisation; human dignity; informed consent; profiling; statistical discrimination

1. Introduction

Applications of artificial intelligence (AI), algorithmic differentiation, and automated decision-making systems (ADMs) are increasingly being used to support and automate decisions about the differentiation of persons in areas as diverse as lending, housing, recruitment, welfare benefits assessments, or judicial decision-making. Such differentiations then affect the availability and distribution of products, services,

positions, opportunities, benefits, or burdens, including those that are essential for personal development and the realisation of autonomy and freedom. Here, applications are used that employ machine learning as an analytical method in data mining, profiling, or predictive analytics, and the results of the analysis are used as models or algorithms in decision-making processes.

The article examines issues that arise when applying a human dignity perspective to algorithmic differentiation and discrimination. The concept of human dignity is not monolithic but includes different dimensions of concretisation (e.g., Mahlmann, 2012; Teo, 2023; von der Pfordten, 2023). Among these dimensions, the most relevant for the following are: the protection against instrumentalisation and the protection of self-determination in shaping one's own life by structuring and pursuing one's own interests, desires, and goals; the protection from false and unjustified degradation or humiliation and from being treated without equal moral worth; and the protection of essential conditions for exercising self-determination. The article discusses the constitutional anchoring of human dignity to gain insights into the impact of algorithmic differentiation on fundamental rights.

2. Algorithmic Differentiation and Discrimination

The causes of bias in the use of algorithms, especially of machine learning as a form of AI, are manifold. They result from human decisions about the database used and the development, deployment, or adaptation of algorithms. The most commonly cited causes of bias include training datasets contaminated with historical inequalities and unequal treatments, missing data that would represent certain groups, the selection of inappropriate labels, measurements, or algorithms, inappropriately decided technical trade-offs, or the application of algorithms in domains for which they have not been trained or optimised (e.g., Mehrabi et al., 2021; Pessach & Shmueli, 2022). Bias can lead to products and services not working equally well for different populations or to algorithmic models used to differentiate persons in decisions about access to and distribution of products, services, positions, or freedoms resulting in unequal treatment of those affected. Another problem is stereotyping in generative AI systems (such as ChatGPT), especially when decisions about humans are based on their output (e.g., in automated analysis or summaries of job applications).

Numerous research and development efforts are directed towards making algorithms less discriminatory or "fairer," especially by adapting and cleaning datasets or modifying algorithms. Mathematical fairness definitions or fairness metrics have been developed to express unequal treatment quantitatively and to optimise and compare systems. However, developers and providers must decide on various trade-offs. These include whether and which fairness definition to use, determining the residual risks that affected persons are exposed to, but also trade-offs between individual goals and metrics, for example, between accuracy in achieving differentiation goals and avoiding discrimination risks (e.g., Mehrabi et al., 2021; Pessach & Shmueli, 2022).

Some fairness metrics are ratios formed from the error rates of "false negatives" and "false positives" and the rates of correctly identified classifications related to certain population characteristics. However, it is still unclear how these error rates or fairness metrics will be dealt with in society, which of the fairness metrics should be used to meet certain societal expectations of justice in specific situations, and whether, how, and what levels of residual discrimination risks will have to be borne by society or in which cases residual risks are not acceptable at all. According to the European Union Agency for Fundamental Rights (FRA), when

using fairness metrics, it can only be decided on a case-by-case basis when there is sufficiently significant discrimination, not by setting an abstract threshold (FRA, 2022, p. 25).

The European Union AI Act (adopted text as of 13 March 2024, AI Act in the following; see European Parliament, 2024) refers to the metrics but does not clarify who sets acceptable risk levels, error rates, or limits. In addition, terms such as "residual risk," "acceptable," and "as far as technically feasible" (Article *9*(5) of the AI Act) or "appropriate level of accuracy" (Article *15*(1) of the AI Act) allow for balancing risk avoidance with cost-effectiveness considerations. This should be seen against the background that costs are incurred for the testing of AI systems and risk avoidance. The AI Act leaves open whether the European Commission, standardisation organisations, developers, providers, users, or actors that certify compliance with the regulation will make the normative decisions about discrimination levels and thus about the social realisation of justice. This can lead to residual risks of discrimination at levels undetermined by society. Individual ADMs and AI systems can have a wide reach, for example, due to market concentration, if one system is used in many companies or administrations, or if used as a component (as general-purpose AI component, foundation model, or "AI as a service") in many other systems. Despite seemingly low error rates, this can lead to systematic discrimination affecting a large proportion of the population.

Residual discrimination risks are confronted with anti-discrimination laws. Bias problems lead to legally defined discrimination if biased algorithms are used in differentiations that result in unjustified unequal treatment when using legally protected characteristics (e.g., gender, ethnicity, religion, disability, age, or sexual identity) or when using seemingly neutral characteristics, procedures, rules, or practices that have a connection to the protected characteristics and then actually make groups with protected characteristics worse off (Hacker, 2018; von Ungern-Sternberg, 2022).

The anti-discrimination law, however, has weaknesses with regard to algorithmic discrimination, because in view of algorithmic, often personalised or individualised differentiations, it can be difficult for affected individuals to perceive unequal treatment compared to others and to provide the legally necessary initial evidence of being in a worse position than comparable other persons. However, these are prerequisites for legal proceedings to be initiated, even if the person or entity accused of discrimination has the burden of proof that they are not discriminating (Orwat, 2020, pp. 72–73; von Ungern-Sternberg, 2022). The already high hurdles for affected individuals, which often prevent a legal discrimination case from being filed, are raised even further.

3. The Understanding of Human Dignity in Constitutional Law

Human dignity is often seen as an abstract concept that can be interpreted in different ways (e.g., Mahlmann, 2008). This is a frequent criticism of its use. However, human dignity is included in many human and fundamental rights documents and has been concretised through implementation in legal systems and jurisprudence (e.g., Mahlmann, 2012; McCrudden, 2008). In the following, reference is made to the decisions of the German Federal Constitutional Court and to the discussions and developments of the constitutional law because human dignity is considered to be largely concretised there.

According to this, human dignity is a fundamental claim to value and respect to which every human being is entitled. Human dignity is inherent in the human being. Everyone possesses it, regardless of their

characteristics, achievements, or social status (BVerfGE [Decision] 87, 209; see Federal Constitutional Court, 1992, para. 107). Above all, it includes the protection of personal individuality, identity, and integrity as well as elementary legal equality (BVerfGE 144, 20; Federal Constitutional Court, 2017, headnote 3a, para. 539). The understanding "is based on a conception of human beings as persons who can make free and self-determined decisions and shape their destiny independently" (BVerfGE 144, 20; Federal Constitutional Court, 2017, para. 539). Here, human dignity is primarily concretised by self-determination in shaping one's own life (von der Pfordten, 2023, pp. 48–50).

To further substantiate human dignity, the so-called "object formula" was developed. According to this formula, it is incompatible with human dignity to make the human being a mere object of state action (BVerfGE 27, 1; Federal Constitutional Court, 1969, para. 33). According to the object formula, the human being may not be treated like a thing, reified, or degraded to a mere object. For further concretisation, the Federal Constitutional Court has developed the "subject formula," according to which it is prohibited to treat individual persons in a way that fundamentally calls into question their subject quality by lacking respect for the value they have for their own sake (Hong, 2019, pp. 418–428, 672–690). Höfling (2021, para. 16) suggests that to determine a violation of human dignity in concrete decision-making situations, it is necessary to consider whether the subject status of a human being is still secured by compensation mechanisms despite the objectification in relationships of subordination and dependency.

Certain forms of discrimination directly constitute a violation of human dignity. This is seen, among other things, in the case of direct discrimination by encroachment on the fundamental rights of freedom everyone is entitled to as set out in Article *3*(3) of the German Basic Law (Herdegen, 2022, para. 120). Höfling (2021) sees an unacceptable violation of human dignity in racial discrimination and similar humiliating unequal treatment (Höfling, 2021, para. 35; see also, in particular, BVerfGE 144, 20; Federal Constitutional Court, 2017, para. 541). Hillgruber (2023, para. 17) emphasises that a violation of human dignity occurs not only when people of a certain "race," skin colour, religion, or gender are regarded as "inferior," but also when people are discriminated against based on a physical or mental disability, especially when there is a threat of exclusion because of their disability. The characteristics mentioned are particularly relevant because, on the one hand, they are immutable and personality-constituting characteristics. On the other hand, they are historically justified as a demarcation from the atrocities of the National Socialist injustice regime (Hong, 2019, p. 407; Lehner, 2013, pp. 226–248).

Human dignity is further concretised through its development into the constitutional general right of personality. This includes, among others, the right to informational self-determination (developed to realise the protection of human dignity and free development of personality), the prohibition of discrimination, but also the right to self-expression (Britz, 2007), which are particularly relevant for the following considerations. Thus, anti-discrimination law not only serves to realise the right to equal treatment and socio-political goals but also the right to free development of personality and the protection of human dignity. In this way, Baer (2009) sees dignity as the promise of recognition of different perceptions of self, all of which deserve equal respect. The right to informational self-determination and the prohibition of discrimination serve, among other things, to prevent inappropriate external images of one's personality. These rights should enable individuals to co-decide what they regard as belonging to and constituting their personality (Britz, 2007, p. 16). The core guarantee of the right of personality is to provide mechanisms that involve individuals in the processes of constituting personality in such a way that they can understand their personality as freely chosen (Britz, 2008, p. 191).

In a series of decisions, the Federal Constitutional Court has specified the right to human dignity and personality and related them to the risks posed by information and communication technologies. In the "microcensus" decision (BVerfGE 27, 1; Federal Constitutional Court, 1969), the court made it clear that it is contrary to human dignity to make human beings mere objects in the state. It is incompatible with human dignity to compulsorily register and catalogue human beings in their entire personality and thus treat them as a thing that is available to an inventory in every respect (para. 33).

The census decision (BVerfGE 65, 1; Federal Constitutional Court, 1983) established the fundamental right to informational self-determination, which serves the right to free development of personality in conjunction with the right to human dignity. It guarantees individuals the authority to decide for themselves about the disclosure and use of their personal data (headnote 1). The right to informational self-determination not only serves to secure the external and internal dimensions of freedom (freedom of action and identity formation) but also to avoid chilling effects that may arise for data subjects due to uncertainties about data processing (Britz, 2010).

In its decision on acoustic surveillance (BVerfGE 109, 279; Federal Constitutional Court, 2004), the court recognises a core area of private life that enjoys absolute protection concerning the inviolability of human dignity. What belongs to the core of private life depends on whether the facts of the case are of a highly personal nature (para. 123). Similarly, sweeping surveillance, in terms of time and location, violates human dignity if it is carried out over an extended period of time and all movements and expressions of the life of the person concerned are recorded and can become the basis of a personality profile (para. 150).

In the decision on the right to be forgotten I (BVerfGE, 152, 152; Federal Constitutional Court, 2019), the court interprets the right to informational self-determination for relationships between private actors in such a way that the right ensures the individual "substantial influence in deciding what information is attributed to their person" (para. 87). The court found that in many life situations, private companies provide the basic services that play a crucial role in shaping public opinion, allocating or denying opportunities, or enabling participation in social or daily life. In many cases, this is done based on extensive data collection and processing, often by companies with market power, where large-scale disclosure of personal data can hardly be avoided in order not to be excluded from the services or opportunities (para. 85). In cases of extensive dependencies or imposition of unacceptable contractual conditions (para. 85) or "where private companies take on a position that is so dominant as to be similar to the state's position…the binding effect of the fundamental right on private actors can ultimately be close, or even equal to, its binding effect on the state" (para. 88).

In the court's ruling on automated data analysis in police work (BVerfG 1 BvR 1547/19; Federal Constitutional Court, 2023), the court emphasises, among other things, that automated data analysis can be used to generate new otherwise inaccessible information, that come close to full profiles of persons affected, by linking existing datasets (para. 69). The court also refers to the discrimination risks of automated data analyses (para. 100), which become less tolerable the more the analyses are capable to produce disadvantages prohibited under Article *3*(3) of the German Basic Law (para. 77). Furthermore, it stresses the importance of the ability to scrutinise algorithms for individual legal protection and administrative oversight in order to identify and correct errors (para. 90). A risk of loss of state oversight is seen particularly in the use of self-learning systems or AI, as these can become detached from the original human programming in the course of the machine learning

process and results become increasingly difficult to scrutinise (para. 100, with reference to the judgement of the European Court of Justice, 2022).

4. Factors of Violation of Human Dignity With Algorithmic Differentiation

4.1. Severe and Structural Discrimination

Algorithmic differentiation can lead to systematic or structural discrimination due to the possibility of residual discrimination risks and the wide reach of the system, which can cover entire populations. Some factors may indicate a violation of human dignity when discrimination is based on "race" or ethnicity, gender, physical and mental disabilities, and the other protected characteristics of Article *3*(3) of the German Basic Law, and when algorithmic differentiation is used in application contexts where there is a strong dependency on the benefits, products, services, or where applications affect particularly vulnerable persons. These groups are considered particularly worthy of protection under constitutional law. Severe discrimination would violate their dignity, as they are systematically degraded. This applies in particular to decisions concerning a dignified life (e.g., in the cases of welfare recipients, refugees, or migrants). Here, discrimination can involve forms of humiliation or degradation, as a person may not be regarded and treated with equal moral worth (e.g., the SyRi and Robodebt scandals; see Teo, 2023, p. 17).

In this respect, algorithmic differentiations that consolidate or expand structural inequalities through negative feedback loops are also problematic. In addition to exacerbating the problem of not being treated as persons of equal moral worth, negative feedback loops can impair the self-determination of those affected by making it difficult or even impossible for them to escape inappropriate classifications or stereotypes on their own. Such negative feedback loops can arise when results from algorithmic differentiation systems that predict the behaviour of persons are recaptured and the systems use this uncorrected data as the basis for further data analysis, inference, or the further development or learning of the algorithms. Examples can be found in predictive policing systems (FRA, 2022; Lum & Isaac, 2016).

4.2. Generalisation and Lack of Individual Justice

Algorithmic differentiation often takes on and alters forms of so-called statistical discrimination (Barocas & Selbst, 2016; Binns et al., 2018). Statistical discrimination is a form of proxy discrimination. Instead of using an elaborate case-by-case examination to determine the actual personality traits or the differentiation target (e.g., the social construct "actual ability to pilot an aircraft"), comparatively easy-to-obtain proxy information (e.g., age in years) is used. This form of differentiation is intended to efficiently overcome an information deficit. Discrimination can occur if the proxies are legally protected characteristics or contain characteristics that correlate with protected characteristics (e.g., Britz, 2008; Hellman, 2008; Schauer, 2018). The proxy information can be derived from empirical studies or, in the case of machine learning, be present in models trained on data.

However, statistical discrimination and generalisation (either by human decision-makers or through the use of algorithms) are already ethically problematic in themselves because group information is applied to individuals and thus acts as quasi-stereotypes and prejudices in decision-making (Gandy, 2010, p. 34). In principle, case-by-case justice is not guaranteed as there is no case-by-case examination and the

individual subject characteristics and the individual situations and contexts are not considered (Britz, 2008). What is often referred to as "prediction" in AI research and practice is not a prediction of an individual's potential behaviour derived from an individual examination. Rather, it is an assignment of individuals to categories, scores, or rankings that are formed statistically or by machine learning and are expected to produce certain outcomes for the sorted individuals in the future.

Moreover, in many cases of algorithmic differentiation, the categories to which individuals are assigned are constructed using data from groups that do not contain the individuals who are actually decided upon (Eckhouse et al., 2019, pp. 198–199). In addition, the categories constructed with AI are usually not comprehensible to those affected or to third parties. In contrast to the use of clearly communicated criteria (e.g., age limits in public administration), such decision criteria and rules evade scientific and public scrutiny and discussion, in particular, whether there is a causal relationship between the criteria and the differentiation goal at all, whether this can be substantiated with evidence or whether the use of certain criteria is socio-politically or morally controversial or undesirable. The main aim of scientific and public scrutiny is to avoid unfounded or spurious correlations (Schauer, 2018), which carry the risk of unjustified degradation.

4.3. Not Treating Persons as Individuals and With Respect

In contrast to the individualisation of decisions about people, statistical discrimination and generalisation treat people as information objects rather than as individuals. The question of when it is morally problematic to treat people not as individuals but only as members of a group (stereotyping) is controversial (e.g., Beeghly, 2018; Lippert-Rasmussen, 2011).

This question often brings to mind the moral considerations of Kant, who demands respect for human beings and the prohibition of instrumentalisation (according to Dillon, 2022, Chapter 2.2; see also Hill, 2014, pp. 315–318). Thus, every human being is required to "acknowledge, in a practical way, the dignity of humanity in every other human being. Hence there rests on him a duty regarding the respect that must be shown to every other human being" (Kant, 1797/2017, p. 225). The respect that people owe each other and that they can demand from other people is respect for their dignity (Kant, 1797/2017, p. 225; according to Schaber, 2016, p. 256; Ulgen, 2017, 2022, pp. 14–15). Respecting the dignity of another person (and of oneself) means treating others "always at the same time as an end, never merely as a means" (Kant, 1785/2012, p. 41).

According to Schaber (2016), who interprets Kant's explanations of the false promise for this purpose, Kant also means that one treats another person merely as a means if one treats the other person in a way the person cannot possibly consent to. This is the case if they have no reason to do so and would not behave rationally if they consented. Respecting the dignity of another person therefore means treating them in such a way that they can reasonably consent (Schaber, 2006, p. 256).

Drawing on Kant, Korsgaard (1996) elaborates the idea that the test for treating another person as a mere means is whether the other person can consent to the way they are treated. In cases of coercion or deception, the other person cannot do so, since in both forms of treatment the other person has been given no chance to choose the end. Hence, treatment is morally wrong if other persons are not able to choose.

She therefore concludes that coercion and deception are, according to Kant's formula of humanity, the most fundamental forms of wrongdoing towards others, the root of all evil (Korsgaard, 1996, pp. 137–140). Schaber (2013, pp. 134–136) adds that deception can be problematic in situations where it impairs the rights of the affected persons to determine essential aspects of their own lives.

Also referring to Kant, Ulgen (2022) develops requirements for treating persons with respect for their inherent dignity with regard to AI. Dignity arises from the autonomy and rational capacity of humans to exercise reasoning, judgements, and choices. Human autonomy is protected if humans "are able to act influenced by reason; if they can identify the motivations prompting their action; or they can change their motivations if they cannot identify with them" (Ulgen, 2022, p. 19). AI systems that diminish the human agency to exercise reasoning, judgement, and choice undermine human dignity (Ulgen, 2022, p. 27). The argument is also relevant to technologically implemented social rules (see Section 4.6).

In summary, these arguments show the importance of having requirements for how persons are treated as well as functioning compensation mechanisms to prevent individuals from being treated as mere objects. This includes, in particular, the opportunity to consent to and influence treatments. The protection of human dignity also includes the protection of choice of ends and the requirement to be informed about choices in such a way that individuals can act in a self-determined manner.

Philosophical approaches to explaining when differentiation is morally wrong come to similar conclusions. These include approaches based on discrimination theory oriented towards human dignity and disregard (Khaitan, 2015, pp. 6–8), even if the term "dignity" is not always used. They consider a differentiation to be wrong if the discriminating person assesses the moral value of the discriminated person wrongly, in particular as lower, or if the discriminating person expresses a wrong assessment, i.e., acts as if the discriminated person has a lower moral value (Thomsen, 2017).

For Hellman (2008), the moral wrong of discrimination is that it degrades a person. Degrading rules or practices express a disregard for the moral equality of those being discriminated against. According to Hellman (2016) it is important to first clarify the meaning of the term "dignity" before using it. Discrimination is wrong from the perspective of human dignity because it results either from the fact that people are not treated with equal worth, i.e., people do not receive the same level of recognition and respect, or from the fact that people are denied rights to which they are entitled (Hellman, 2016, pp. 943–946). Such rights can include the right to self-determination over essential aspects of life and the right to receive a justification.

Eidelson (2015) conceives the wrong of discrimination as a failure to treat a person correctly as an individual. The error lies in not seeing the person as the (partial) result of their past efforts at self-creation and as an autonomous agent whose future decisions they can make for themselves. About the problem that statistical discrimination and generalisation do not treat persons as individuals, he therefore calls for an understanding that persons are treated as individuals if—and only if—(a) the differentiating person X gives appropriate weight to the evidence of how the affected person Y has exercised autonomy in shaping his or her life, provided that this evidence is reasonably available and relevant to the decision at hand; and (b) in addition, X's judgements, when they relate to the choices of person Y, must not be made in such a way as to disparage Y's capacity to make those choice decisions as an autonomous agent (Eidelson, 2015, pp. 144, 227).

Although questions remain about the scope and types of adequate and relevant evidence and about obligations to provide it, it can be deduced that the object of information and decision-making should be the self-determined personality development of the persons concerned, in particular their possibilities of self-perception, self-determination, and self-expression. However, a dilemma must be avoided. If personal data on self-determination is to be collected to solve the problem of generalisation and to better respect persons as individuals, this can only be done by having the data and profiles controlled by the persons affected themselves to avoid a violation of the right to informational self-determination. In addition, it may be necessary not to rely solely on automated data collection and analysis but to involve human decision-makers to collect sufficient additional information and to make situation- and person-specific judgements that require a high degree of situational balancing when evaluating the information provided.

4.4. Automated Decision-Making

According to Citron (2008) and Kaminski (2019), automated decision-making based on generalisations are ethically problematic because no other information about the persons concerned is processed apart from the generalisation information. If persons are merely assigned to algorithmically formed categories, scores, or rankings, persons are no longer treated as individuals. If automated decisions no longer allow persons to express their individuality, this violates their dignity and turns people into objects based on a few characteristics instead of treating them as whole persons. Both the exercise of human discretion and individual procedural rights (of appeal, correction, etc.) are necessary not only to avoid error but also to properly recognise and respect individuality (Citron, 2008, p. 1304; Kaminski, 2019, pp. 1541–1545). Furthermore, human discretion is necessary when human decision-makers must also be able to consider mitigating circumstances that the algorithm cannot, as well as when there are indeterminate terms in the decision rules that require the human decision-maker to make trade-offs between conflicting interests (Citron, 2008, p. 1304).

One of the justifications for regulating automated decision-making is the protection of human dignity. This refers to Article 22 of the General Data Protection Regulation (GDPR; European Parliament and Council of the European Union, 2016) and its predecessor, Article 15 of the Data Protection Directive 95/46/EC (European Parliament and Council of the European Union, 1995). According to Dammann and Simitis (1997), the prohibition of automated decision-making of Article 15 of the Data Protection Directive was intended to prevent data subjects from being treated in personality assessments only by computer and based on stored data. This would ignore the individuality of the person and degrade the person to a mere object of computer operations (Dammann & Simitis, 1997, pp. 218–219; see also Jones, 2017; Kaminski, 2019; Martini, 2021, para. 8; Scholz, 2019, para. 3).

According to Martini and Nink (2017), the subject quality of a person is not necessarily disregarded by the fact that personal data alone are the object of an algorithmic analysis. In the case of automated (administrative) decisions, the quality of the subject is only affected when algorithmic procedures impose adverse consequences on the person concerned without allowing them to defend themselves against the decision in an appropriate way. To protect informational self-determination, the legal practice relies on the procedure of (a) informing about the automated decision, (b) communicating and explaining the essential reasons for the decision upon request, (c) allowing the data subject to assert their own point of view in order to obtain a review and re-evaluation if necessary (Martini & Nink, 2017, p. 7).

However, some deficits in the actual regulation of automated decision-making raise doubts as to whether it can still serve to protect human dignity. The so-called "prohibition" is provided with extensive exceptions, particularly if the automated decision-making serves to conclude or fulfil a contract, is required by law to be permissible, or explicit consent is given. Although the regulation provides that the operator of an automated decision must inform about the existence of an automated decision and the so-called logic involved, it is still unclear what the content of this information obligation is, in particular, whether and how information about decision criteria or possible discrimination risks must be provided (Orwat, 2020, pp. 77–82).

Moreover, the "prohibition" is often interpreted only as a right of intervention for the persons affected in justified individual cases (Martini & Nink, 2017, p. 4). In this context, the persons affected must first have knowledge of the automated decision-making and its effects, and justify their desire for human intervention and an explanation of the logic involved. As this can be very burdensome, it may have a chilling effect if individuals perceive it as an unreasonably high hurdle to avail themselves of the regulation. Persons affected may be deterred from exercising the rights to which they are entitled and which were established to protect their dignity.

4.5. Emergence of New Knowledge and Comprehensive and Meaningful Profiles

The possibilities of data aggregation, data reuse, data combination and inference, de-anonymisation and re-identification of individuals, categorisation, ranking, assessment, and individual or group profiling of individuals have greatly increased with AI (e.g., FRA, 2020; Smuha, 2021; Yeung, 2019). Some AI systems have been developed to make automated inferences about identity, personality-constituting traits, and other sensitive information such as emotions, character traits, mental states, or political orientations (e.g., Kosinski, 2021; Matz et al., 2023). AI-based biometric and psychometric evaluations (e.g., emotional AI) can be used for targeting (e.g., in marketing), risk assessment (e.g., in applicant selection or calculating the probability of dropping out of university or defaulting on a loan), and behavioural control (Valcke et al., 2021). With other AI methods, researchers strive to predict future events in the lives of persons, including personality nuances and the time of death (Savcisens et al., 2023).

The systems are often based on a reduction of personality to quantifiable measures and classes that attempt to map the personality traits relevant to a differentiation goal. Further critical issues are the standardisation of personality (Köchling et al., 2021) or the pseudo-scientific approach (Sloane et al., 2022). Even if the scope and types of application of such systems are still little known in practice, this illustrates that AI can generate and use sweeping and meaningful profiles that are suitable for imposing an (almost) complete external image on a person, with personality-constituting characteristics and even without the valid consent of the person concerned (see Section 4.7).

All in all, this leads to a further detachment of the data representation by the operators from the possibilities of controlling the self-representation by the persons concerned (cf. Teo, 2023). The options of agency and the identity of those affected are then solely determined externally, including how they (have to) see themselves (as normal, healthy, conforming to rules, etc.). The abilities to develop and shape their life in a self-determined way are then eliminated. According to the standards outlined above, this can be a violation of (informational) self-determination and human dignity. Such AI applications can also have chilling effects and thus lead to self-restriction of the free development of personality as a form of violation of dignity.

4.6. Structural Dominance

In the relationship between the state and persons affected (e.g., citizens or migrants), the structural dominance of the state must be assumed as a matter of principle, because there are usually situations with a monopoly on the use of force, a lack of options for evasion, subjection, non-negotiability, and the complete, unilaterally determined binding nature of the rules. A number of factors can also increase the structural dominance of users of AI systems (e.g., providers, employers, banks) over those affected in private relationships (e.g., customers, applicants, credit seekers).

Firstly, social rules are increasingly being implemented in software or algorithms in both the public and private sectors. For technical implementation, the rules must be written in programming languages or generated in the form of algorithms or models through machine learning. With these forms of specification, however, the scope for interpretation and discretion of social rules for human decision-makers is also lost, which is often required so that social rules can be applied to many, sometimes unpredictable, situations. If rules are enforced fully automatically, such as in fully automated decision-making, deviation from the rules is normally technically prevented. Also, in algorithmic choice architectures (nudging), the space of choices is technically predetermined and often limited. The rules are then often established unilaterally by those developing and applying them, and the affected persons' possibilities for negotiation, contesting, influencing, and correction are reduced or eliminated, which can increase structural dominance. If the scope for interpretation, discretion, and choice is reduced, the possibilities for action, autonomy, and the opportunities for self-enforcement of autonomy by those affected are also reduced (Deutscher Ethikrat, 2023, pp. 120–137; Teo, 2023, pp. 27–31; Ulgen, 2022). Secondly, if the private users of AI systems are also those who operate platforms (and in some cases also those who develop AI systems), strong network effects of the platforms can reduce the opportunities for evasion and increase users' dependency on the applications. Thirdly, as just shown, some AI systems can determine sensitive personality traits such as mental states, character traits, and emotions even from seemingly "trivial" data such as communication in social networks. In this way, reliance on a product, service, or position can be better determined and exploited (e.g., Härtel, 2019), as can human weaknesses, especially when the systems are used for "dark patterns" or other forms of manipulation (e.g., Ulgen, 2022, pp. 22–24).

Developments that increase the structural dominance of the users of AI systems over those affected can tend to restrict human dignity and the free development of personality because those affected are restricted in their ability to shape their own lives. This can happen in private relationships by way of disrupting contractual parity and thus reducing the possibilities for those affected to assert their self-determination themselves. For even if freedom of contract applies, i.e., everyone has the freedom to determine with whom and under what conditions contracts are entered into, it must be ensured that also the conditions of free self-determination are actually given (BVerfGE 81, 242; Federal Constitutional Court, 1990, para. 47).

4.7. Absence of the Possibility of Valid Consent

The instrument of consent as a compensation mechanism for treating persons as mere objects can transform morally impermissible treatments into permissible ones. However, it can only achieve this moral transformation if certain preconditions are met. These include that the affected persons consent voluntarily and have choices to do so, that they are sufficiently informed, i.e., understand the data processing,

decision-making, and the consequences thereof, and that they have the necessary decision-making skills (Bullock, 2018). On the other hand, from a philosophical point of view, it is also doubted that consent can transform treatment as a mere object into morally permissible treatment if the treatment already violates the duty to treat other people with respect (Fahmy, 2023). This is likely to be the case, for example, with serious algorithmic discrimination or differentiations based on profiles with (almost) complete recording and determination of personality without control by the persons affected.

In the field of data protection, the problems of informed consent have long been recognised. The effectiveness and meaningfulness are increasingly limited by non-negotiable, long, incomprehensible data protection declarations or privacy policies formulated in legalese, the increasing collection of data that is based on so-called legitimate interests without the need for consent (Article *6*(1)f GDPR; European Parliament and Council of the European Union, 2016), strong network effects and thus the tying of customers and users to systems or platforms, which reduces voluntariness as well as interface designs that entice consent to data collection. Those affected often lack knowledge about the necessity and legal possibilities of informed consent. They can hardly assess the actual consequences of consent regarding potentially detrimental, sometimes long-term treatments, which can also arise from data accumulation that is difficult to trace or further data processing and transfer that can no longer be assessed. Furthermore, the decision criteria of complex AI algorithms may be incomprehensible or unknown, particularly in the case of self-learning or adaptive systems. Similarly, AI-based reasoning can generate new knowledge from existing personal data, even from anonymised data, and even for individuals or groups not involved in the original consent. It can then be assumed that the data subjects can no longer sufficiently recognise what they are consenting to (e.g., Orwat, 2020, pp. 71–72). Due to these factors, established types of informed consent becomes increasingly useless as a legitimisation of the treatment of people as mere objects and as an instrument of self-determination.

Zarsky (2013) explains in more detail that in order to protect human dignity, there must be an understanding of the inner workings of automated data analyses, as without this understanding the results can still appear arbitrary and wrong. The problem is that existing (legal) transparency requirements are at best sufficient to provide information about the correlations or classifications a person might fall into. Instead, the automated prediction process must also be interpretable, i.e., the selection process must be explainable. The protection of human dignity therefore requires that causations and not merely correlations must be ascertainable for those affected before conclusions and measures are taken (Zarsky, 2013, p. 1548).

5. Conclusions

The human dignity perspective on algorithmic differentiation and discrimination can help determine how humans or machines should treat other humans as individuals and with respect. This perspective can complement work on the (relative) fairness of algorithmic differentiation and the technical mitigation of discrimination risks. The perspective of human dignity goes beyond efforts to de-bias datasets and algorithms. It leads to the question of how algorithm-based decision-making applications should be shaped and what information bases and forms of communication should be used. It also helps differentiate more clearly between different application situations, for example, for which applications and purposes it should be able or allowed to use the new capabilities of AI. Moreover, the perspective provides justifications for when AI and ADMs should not be used due to the possible restriction or violation of human dignity.

For example, the applicability of AI and ADMs is questionable if decision-making situations are unavoidable for those affected and they have de facto no possibility to influence decisions, if self-determination is severely restricted, or if serious degradation may occur. It also points out that even the ideal case of "accurate" profiles or categories as the basis of decisions is problematic if overpowering external profiles suppress the informational self-determination of those concerned.

The use of AI and algorithmic differentiation can lead to a violation or restriction of human dignity. Problematic factors include: (a) the use of generalisations and disregard of personality in decisions on unequal treatment using variables (proxies) that have no comprehensible connection to the differentiation goal, use morally dubious connections, lack rational justification, and cannot be contested by the persons concerned; (b) the reach of systems with residual risks of systematic and structural discrimination, which, among other things, expose some persons to higher risks of discrimination, do not guarantee equal protection for all, and thus treat them as persons with inferior moral worth; (c) the increasingly inadequate established types of informed consent, which has a particularly drastic effect with AI systems whose decision-making criteria and far-reaching consequences are no longer comprehensible to those affected; (d) the inadequate regulation of (fully) automated decisions, the role of the human decision-makers involved in them, and informed consent to it; (e) the loss of control over the generation and use of sweeping and meaningful personal and group profiles by those affected, leaving no scope for self-determination over the external image created; and (f) the increasing structural dominance of the state and private companies through increasing technical enforcement of social rules, market concentration, specific capabilities of AI to generate new knowledge about persons from existent data and to detect and exploit dependencies or other human weaknesses and thus situations with distorted contractual parity, strong dependencies, limited options for action, contestation, influence, avoidance, and the resulting inevitability. As a result, in situations where these factors have an impact, either alone or in combination, the protection of human dignity can no longer be guaranteed. These impacts are all the more severe the more essential the products and services are for the self-determined shaping of life and identity, or for a dignified existence of people with special needs and vulnerabilities (e.g., access to education, employment, health and welfare services, finance, housing, or asylum).

It is important to further clarify when human dignity and the development of personality are specifically restricted or violated and how they can be protected. A context-specific approach is necessary as different products, services, or treatments have varying degrees of relevance to the preservation of dignity. The prohibitions of certain AI applications in the European Union AI Act is justified, among other things, by the aim of protecting human dignity (e.g., Recital 28 and 31 AI Act; European Parliament 2024). The prohibitions include certain social scoring systems, subliminal, manipulative, or deceptive techniques, systems that exploit vulnerabilities (due to age, disability, social or economic situation), systems that infer emotions in the workplace or in education, biometric categorisation systems that use or infer sensitive characteristics (e.g., political opinions, religious or philosophical beliefs, sexual orientation, race), with many exceptions, certain remote biometric identification systems for law enforcement, and other applications. However, the following considerations may help interpret and develop the AI Act, the regulatory framework in general, and in designing AI and ADM applications. The issues include:

1. Which personal and group profiles are so comprehensive or constitutive of one's personality that the image of the personality must be described as externally determined, as no longer freely chosen, and

the self-determination of essential aspects of the shaping of life as undermined (e.g., if the boundary to social scoring is exceeded, if an external profile unduly restricts access to services and products essential for personality development or if it is completely beyond the control of those affected)?

2. Under what conditions does serious, systematic, or structural discrimination actually exist or is to be expected, and in what areas can residual discrimination risks not be tolerated as this would lead to a violation of dignity (e.g., in areas where there are no or limited possibilities to influence or evade decisions that belong to the core of a self-determined shaping of life, for persons with special vulnerabilities or large dependencies, or where decision-making could result in a massive, unjustified, serious degradation of persons)?
3. To what extent and in what form should the personality of those affected be respected in algorithm-based decisions, and what form of justification for decisions must those affected receive (e.g., by explaining and justifying the use of causal relationships between the criteria used and the differentiation goals, by providing all information necessary for self-determination, by preventing trade secret interests from taking precedence over information claims based on human dignity, or by involving human decision-makers who ensure consideration of specific personal information and situational balancing)?
4. What options and capabilities must those affected have to influence decisions and their personality profile, and what should communicative processes look like in this respect (e.g., requirements for the explainability or comprehensibility of AI and ADMs to provide an adequate information basis for contesting and influencing decisions by those affected)? Where are the capabilities of those affected inappropriate or limited and other institutional actors should become active (e.g., through collective redress)?
5. How is it possible not only to (re)strengthen the human dignity and personality development of those directly affected but also to ensure the protection of groups or third parties who are unaware that they are affected?

Acknowledgments

The author would like to thank the three anonymous reviewers as well as his colleagues Harald König, Philipp Frey, Reinhard Heil, Michael Schmidt, Sylke Wintzer, and the editor of the thematic issue for their valuable feedback.

Conflict of Interests

The author declares no conflict of interest.

References

Baer, S. (2009). Dignity, liberty, equality: A fundamental rights triangle of constitutionalism. *University of Toronto Law Journal*, *59*(4), 417–468.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.

Beeghly, E. (2018). Failing to treat persons as individuals. *Ergo: An Open Access Journal of Philosophy*, *5*(26), 687–711.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). *'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions* [Paper presentation]. 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada.

Britz, G. (2007). *Freie Entfaltung durch Selbstdarstellung. Eine Rekonstruktion des allgemeinen Persönlichkeitsrechts aus Art. 2 I GG*. Mohr Siebeck.

Britz, G. (2008). *Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung*. Mohr Siebeck.

Britz, G. (2010). Informationelle Selbstbestimmung zwischen rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts. In W. Hoffmann-Riem (Ed.), *Offene Rechtswissenschaft* (pp. 561–596). Mohr Siebeck.

Bullock, E. C. (2018). Valid consent. In P. Schaber & A. Müller (Eds.), *The Routledge handbook of the ethics of consent* (pp. 85–94). Routledge.

Federal Constitutional Court. (1969). *Mikrozensus: Beschluß des Ersten Senats vom 16. Juli 1969* (1 BvL 19/63). http://www.servat.unibe.ch/dfr/bv027001.html

Federal Constitutional Court. (1983). *Volkszählung: Judgment of the First Senate of 15 December 1983* (1 BvR 209/83). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/1983/12/rs19831215_1bvr020983en.html

Federal Constitutional Court. (1990). *Handelsvertreter: Beschluß des Ersten Senats vom 7. Februar 1990* (1 BvR 26/84). https://www.servat.unibe.ch/dfr/bv081242.html

Federal Constitutional Court. (1992). *Tanz der Teufel: Beschluß des Ersten Senats vom 20. Oktober 1992* (1 BvR 698/89). https://www.servat.unibe.ch/dfr/bv087209.html

Federal Constitutional Court. (2004). *Großer Lauschangriff: Order of the First Senate of 3 March 2004* (1 BvR 2378/98). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2004/03/rs20040303_1bvr237898en.html

Federal Constitutional Court. (2017). *NPD-Verbotsverfahren: Judgment of the Second Senate of 17 January 2017* (2 BvB 1/13). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2017/01/bs20170117_2bvb000113en.html

Federal Constitutional Court. (2019). *Right to be forgotten I: Order of the First Senate of 6 November 2019* (1 BvR 16/13). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2019/11/rs20191106_1bvr001613en.html

Federal Constitutional Court. (2023). *Automated data analysis: Judgment of the First Senate of 16 February 2023* (1 BvR 1547/19). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2023/02/rs20230216_1bvr154719en.html

Citron, D. K. (2008). Technological due process. *Washington University Law Review*, *85*(6), 1249–1313.

European Court of Justice. (2022). *Ligue des droits humains ASBL v Conseil des ministres* (Case C-817/19). https://curia.europa.eu/juris/liste.jsf?lgrec=fr&td=%3BALL&language=en&num=C-817/19&jur=C

European Parliament and Council of the European Union. (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive). *Official Journal of the European Communities*, L 281/31. http://data.europa.eu/eli/dir/1995/46/oj

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L 119/1. http://data.europa.eu/eli/reg/2016/679/2016-05-04

European Parliament (2024). *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* (COM(2021)0206—C9-0146/2021—2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

Dammann, U., & Simitis, S. (1997). *EG-Datenschutzrichtlinie: Kommentar*. Nomos.

Deutscher Ethikrat. (2023). *Mensch und Maschine—Herausforderungen durch Künstliche Intelligenz*.

Dillon, R. S. (2022). Respect. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2022/entries/respect

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, *46*(2), 185–209.

Eidelson, B. (2015). *Discrimination and disrespect*. Oxford University Press.

Fahmy, M. S. (2023). Never merely as a means: Rethinking the role and relevance of consent. *Kantian Review*, *28*(1), 41–62.

FRA. (2020). *Getting the future right—Artificial intelligence and fundamental rights*. European Union Agency for Fundamental Rights.

FRA. (2022). *Bias in algorithms—Artificial intelligence and discrimination*. European Union Agency for Fundamental Rights.

Gandy, O. H., Jr. (2010). Engaging rational discrimination: Exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology*, *12*(1), 1–14.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(4), 1143–1185.

Härtel, I. (2019). Digitalisierung im Lichte des Verfassungsrechts—Algorithmen, Predictive Policing, autonomes Fahren. *Landes- und Kommunalverwaltung*, *29*(2), 49–60.

Hellman, D. (2008). *When is discrimination wrong?* Harvard University Press.

Hellman, D. (2016). Two concepts of discrimination. *Virginia Law Review*, *102*(4), 895–952.

Herdegen, M. (2022). Art. 1 Abs. GG (Schutz der Menschenwürde). In T. Maunz & G. Dürig (Eds.), *Grundgesetz-Kommentar*. Beck.

Hill, T. E., Jr. (2014). In defence of human dignity: Comments on Kant and Rosen. In C. McCrudden (Ed.), *Understanding human dignity* (pp. 313–325). Oxford University Press.

Hillgruber, C. (2023). GG Art. 1 (Schutz der Menschenwürde). In V. Epping & C. Hillgruber (Eds.), *BeckOK (Online-Kommentar) Grundgesetz*.

Höfling, W. (2021). Art. 1 GG Schutz der Menschenwürde, Menschenrechte, Grundrechtsbindung. In M. Sachs (Ed.), *Grundgesetz: Kommentar* (pp. 70–102). Beck.

Hong, M. (2019). *Der Menschenwürdegehalt der Grundrechte. Grundfragen, Entstehung und Rechtsprechung*. Mohr Siebeck.

Jones, M. L. (2017). The right to a human in the loop: Political constructions of computer automation and personhood. *Social Studies of Science*, *47*(2), 216–239.

Kaminski, M. E. (2019). Binary governance: Lessons from the GDPR's approach to algorithmic accountability. *Southern California Law Review*, *92*(6), 1529–1616.

Kant, I. (2012). *Groundwork of the metaphysics of morals—Revised edition*. Cambridge University Press. (Original work published 1785)

Kant, I. (2017). *The metaphysics of morals*. Cambridge University Press. (Original work published 1797)

Khaitan, T. (2015). *A theory of discrimination law*. Oxford University Press.

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2021). Highly accurate, but still discriminatory. *Business & Information Systems Engineering*, *63*(1), 39–54.

Korsgaard, C. M. (1996). *Creating the kingdom of ends*. Cambridge University Press.

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, *11*(1), Article 100. https://doi.org/10.1038/s41598-020-79310-1

Lehner, R. (2013). *Zivilrechtlicher Diskriminierungsschutz und Grundrechte. Auch eine grundrechtliche Betrachtung des 3. und 4. Abschnittes des Allgemeinen Gleichbehandlungsgesetzes (§§19-23 AGG)*. Mohr Siebeck.

Lippert-Rasmussen, K. (2011). "We are all different": Statistical discrimination and the right to be treated as an individual. *The Journal of Ethics*, *15*(1), 47–59.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19.

Mahlmann, M. (2008). *Elemente einer ethischen Grundrechtstheorie*. Nomos.

Mahlmann, M. (2012). Human dignity and autonomy in modern constitutional orders. In M. Rosenfeld & A. Sajó (Eds.), *The Oxford handbook of comparative constitutional law* (pp. 1–26). Oxford University Press.

Martini, M. (2021). DS-GVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling. In B. P. Paal & D. A. Pauly (Eds.), *Beck'sche Kompakt-Kommentare. Datenschutz-Grundverordnung, Bundesdatenschutzgesetz* (3rd ed.). Beck.

Martini, M., & Nink, D. (2017). Wenn Maschinen entscheiden…—Vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz. *Neue Zeitschrift für Verwaltungsrecht*, *36*(10), 1–14.

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), Article 5705. https://doi.org/10.1038/s41598-023-32484-w

McCrudden, C. (2008). Human dignity and judicial interpretation of human rights. *European Journal of International Law*, *19*(4), 655–724.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Orwat, C. (2020). *Risks of discrimination through the use of algorithms*. Federal Anti-Discrimination Agency.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1–44.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*, *4*(1), 43–56.

Schaber, P. (2013). *Instrumentalisierung und Menschenwürde* (2nd ed.). Mentis.

Schaber, P. (2016). Menschenwürde. In A. Goppel, C. Mieth, & C. Neuhäuser (Eds.), *Handbuch Gerechtigkeit* (pp. 256–262). J. B. Metzler.

Schauer, F. (2018). Statistical (and non-statistical) discrimination. In K. Lippert-Rasmussen (Ed.), *The Routledge handbook of the ethics of discrimination* (pp. 42–53). Routlegde.

Scholz, P. (2019). DSGVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling. In S. Simitis, G. Hornung, & I. Spiecker genannt Döhmann (Eds.), *Datenschutzrecht. DSGVO mit BDSG*. Nomos.

Sloane, M., Moss, E., & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, *3*(2), Article 100425. https://doi.org/10.1016/j.patter.2021.100425

Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3), 1–32.

Teo, S. A. (2023). Human dignity and AI: Mapping the contours and utility of human dignity in addressing challenges presented by AI. *Law, Innovation and Technology*, *15*(1), 1–39.

Thomsen, F. K. (2017). Discrimination. In W. R. Thompson (Ed.), *Oxford research encyclopedia of politics*. Oxford University Press.

Ulgen, O. (2017). Kantian ethics in the age of artificial intelligence and robotics. *Questions of International Law*, *43*, 59–83.

Ulgen, O. (2022). AI and the crisis of the self: Protecting human dignity as status and respectful treatment. In A. J. Hampton & J. A. DeFalco (Eds.), *The frontlines of artificial intelligence ethics: Human-centric perspectives on technology's advance* (pp. 9–33). Routledge.

Valcke, P., Clifford, D., & Dessers, V. K. (2021). Constitutional challenges in the emotional AI era. In H.-W. Micklitz, O. Pollicino, A. Reichman, A. Simoncini, G. Sartor, & G. De Gregorio (Eds.), *Constitutional challenges in the algorithmic society* (pp. 57–77). Cambridge University Press.

von der Pfordten, D. (2023). *Menschenwürde* (2nd ed.). Beck.

von Ungern-Sternberg, A. (2022). Discriminatory AI and the law—Legal standards for algorithmic profiling. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives* (pp. 252–277). Cambridge University Press.

Yeung, K. (2019). *Responsibility and AI. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. Council of Europe.

Zarsky, T. (2013). Transparent predictions. *University of Illinois Law Review*, *2013*(4), 1503–1569.

About the Author

Carsten Orwat (PhD) is a senior researcher at the Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology. Since 2000, he has worked on numerous technology assessment projects on information and communication technologies. His research also focuses on the governance and regulation of technology. Currently, he is researching the social consequences of artificial intelligence, algorithmic discrimination, and systemic risks.

Social Inclusion
2024 • Volume 12 • Article 7471
https://doi.org/10.17645/si.v12.7471

cogitatio

ARTICLE

Open Access Journal

The Artificial Recruiter: Risks of Discrimination in Employers' Use of AI and Automated Decision-Making

Stefan Larsson [1], James Merricks White [1], and Claire Ingram Bogusz [2]

[1] Department for Technology and Society, Lund University, Sweden
[2] Department of Informatics and Media, Uppsala University, Sweden

Correspondence: Stefan Larsson (stefan.larsson@lth.lu.se)

Submitted: 31 July 2023 **Accepted:** 26 February 2024 **Published:** 18 April 2024

Issue: This article is part of the issue "Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination" edited by Derya Ozkul (University of Warwick), fully open access at https://doi.org/10.17645/si.i236

Abstract

Extant literature points to how the risk of discrimination is intrinsic to AI systems owing to the dependence on training data and the difficulty of post hoc algorithmic auditing. Transparency and auditability limitations are problematic both for companies' prevention efforts and for government oversight, both in terms of how artificial intelligence (AI) systems function and how large-scale digital platforms support recruitment processes. This article explores the risks and users' understandings of discrimination when using AI and automated decision-making (ADM) in worker recruitment. We rely on data in the form of 110 completed questionnaires with representatives from 10 of the 50 largest recruitment agencies in Sweden and representatives from 100 Swedish companies with more than 100 employees ("major employers"). In this study, we made use of an open definition of AI to accommodate differences in knowledge and opinion around how AI and ADM are understood by the respondents. The study shows a significant difference between direct and indirect AI and ADM use, which has implications for recruiters' awareness of the potential for bias or discrimination in recruitment. All of those surveyed made use of large digital platforms like Facebook and LinkedIn for their recruitment, leading to concerns around transparency and accountability—not least because most respondents did not explicitly consider this to be AI or ADM use. We discuss the implications of direct and indirect use in recruitment in Sweden, primarily in terms of transparency and the allocation of accountability for bias and discrimination during recruitment processes.

Keywords

ADM and risks of discrimination; AI and accountability; AI and risks of discrimination; AI and transparency; artificial intelligence; automated decision-making; discrimination in recruitment; indirect AI use; platforms and discrimination

1. Introduction

Perhaps the most talked about example of discrimination in AI-supported recruitment is Amazon's now-defunct résumé scanning tool that employed automated decision-making (ADM). Developed between 2014 and 2018, Amazon trained the tool on the credentials of previously recruited candidates to effectively identify and rank qualified applicants. However, the system—which largely looked for patterns—began downgrading résumés from female candidates. Although applicants' gender was not explicitly given, the AI system used indirect markers, such as "captain of the women's chess club," as proxies. After discovering this bias, Amazon's engineers tried to fix the problem by directing the system to treat these terms in a "neutral" way. The company eventually abandoned its efforts: They could not prevent the tool from discriminating based on gender and faced ongoing public criticism (cf. Myers West et al., 2019). This highlights one AI development challenge: the reliance on something already established on which to train (Larsson et al., 2023b).

In the above example, prevailing inequality in the tech industry has been argued to explain why Amazon had such a hard time developing a recruitment tool that did not reproduce that bias (Ajunwa, 2023). The problem thus lies partly in how markers for gender, ethnicity, and age are implicit in the recruitment environments, and partly that AI systems, due to their training data, risk reflecting historical or present biases instead of employers' stated requirements for professional skills. It is certainly not new knowledge that language can be gendered, i.e., that different words can have culturally established male or female associations that, for example, affect how job advertisements are perceived (cf. Gaucher et al., 2011). Through machine learning, though, these associations are often captured and hidden in an AI model (Leavy, 2018), mirroring stereotypes (cf. Larsson et al., 2023b). While well-documented in gendered language, proxies for race have also been seen to influence search engine algorithms (Noble, 2018).

It is not possible to understand a general trend through examination of one exemplar case, like the Amazon résumé scanning tool. Research shows that AI and ADM are increasingly used in worker recruitment, specifically through LinkedIn (Ajunwa & Greene, 2019; Laukkarinen, 2023; Ruparel et al., 2020). However, this does not reveal how recruiters perceive their own reliance on such tools or their underlying data. To better understand the degree to which recruiters are aware when they are using AI, and the implications of this use, we turned to collecting data about how AI and ADM are used in Sweden. Considerable public attention and academic research have been paid to the use of algorithms in the gig economy (e.g., Jones, 2021; Woodcock & Graham, 2020) including in the Nordic region (Ilsøe & Söderqvist, 2022; Newlands, 2022), as well as broader fears of mass unemployment due to automation and robotisation (e.g., Pasquale, 2020). But these are not the only ways that AI and ADM are affecting working life. As an extension of widespread and ongoing social processes of digitisation and datafication (see Thylstrup, 2018; van Dijck, 2014), these tools have also found their way into regular practices of hiring and firing, and organising and monitoring workers (Lomborg, 2022) in a near-infrastructural way (Plantin et al., 2016). This platformisation (van Dijck et al., 2018) of anything automated in recruitment leads to a question of whether AI and ADM are directly used in-house by the professionals, or if it is indirectly used via digital services and platforms like LinkedIn (cf. Komljenovic, 2019). Recent research has attended to expectations, perceptions, and outcomes of hiring algorithms (Dencik & Stevens, 2021; Zhang & Yencha, 2022) and their inherent biases and developer attempts to mitigate them (Kelly-Lyth, 2021; Sánchez-Monedero et al., 2020). The mentioned distinction between direct and indirect uses of AI and ADM is relevant in terms of awareness of risks of discrimination needed for assuring accountability for preventive measures.

The heterogeneous deployment of AI and ADM in the workplace is a significant challenge for regulators. To the extent this deployment is dependent on large external platforms' services, the recently implemented Digital Services Act (DSA; see European Commission, 2022) is of interest, as it imposes transparency obligations on the largest tech firms and their services (including LinkedIn and Facebook). Furthermore, given the classification of employment and worker management as high-risk areas in the EU's forthcoming AI legislation (Veale & Zuiderveen Borgesius, 2021; on the provisional trilogue agreement see also European Commission, 2023), the risk for bias and discrimination in this area is of particular importance in highly digitised EU economies, such as Sweden. Lastly, the purpose of the Swedish Discrimination Act (SFS 2008:567) is to counter discrimination and in other ways promote equal rights and opportunities regardless of the basis of discrimination. Within working life and education, the law also requires employers and education providers to work continuously to promote equal rights and opportunities. There is, however, a challenge of how to interpret established legislation on anti-discrimination in the light of AI and ADM practices (cf. Wachter et al., 2021).

Given the risk of discrimination that seems inherent in the use of these tools, methods, and services, this study's research objective is to explore to what extent and in what contexts employers use AI and ADM for worker recruitment. It particularly engages with questions related to individual and organisational risk awareness, and what this implicates in terms of transparency and accountability for those risks. In what follows, we give some theoretical background to our study (Section 2), before describing our data and methods (Section 3). We then present and analyse the results of our questionnaires with representatives from 10 major recruitment agencies and 100 private sector companies with more than 100 employees (Section 4). We discuss these results in light of the research objective, as well as its implications for AI, ADM, and risk and discrimination (Section 5).

2. Theoretical Background

2.1. Definitional Difficulties

Coined in the 1950s, "artificial intelligence" has proven an elusive concept, both as a field of research (Hagendorff & Wezel, 2020) and in regulatory efforts by European legislators (European Commission, 2021). AI has, for instance, been defined narrowly to include only algorithms that exhibit human-like intelligence—including only algorithms that are self-referential and thus teach themselves (Berente et al., 2021)—and broadly as any kind of automated algorithm, including rule-based systems like decision trees (as per the European Commission's proposal for the AI Act in 2021). According to legal experts Gasser and Almeida (2017, p. 59), "AI is not a single technology, but rather a set of technologies and sub-disciplines ranging from areas such as speech recognition and computer vision to attention and memory, to name just a few." Although the field has developed a great deal in recent years, including paying attention to large language models and what has come to be referred to as "generative AI" or "general purpose AI," there is still much heterogeneity. This is partly a definitional problem, because processes can be automated and use AI in different ways, and a problem with the systems' lack of transparency. This provokes both an object of scrutiny and a methodological question: How are AI-associated risks perceived and mitigated when this concept is blurry and how can we study these risks through questionnaires without defining the concept in detail, respectively? In the following sections, we describe our broad perspective on AI and ADM and use it to explain our approach to these challenges.

Both to avoid getting bogged down in definitional issues and to explore AI in its broadest sense, we treated the definition of AI (and its implicit ambiguities) as part of the research inquiry. Terminological ambiguity around the concept of AI has implications for the questions addressed by this study. We see two challenges here: one methodological and one analytical. The methodological challenge is that we cannot expect our respondents to have a uniform understanding of what constitutes or does not constitute AI. The analytical challenge stems from the wide range of technologies and methods that fall under the umbrella terms "artificial intelligence" and "automated decision-making," and the difficulty of knowing when and where an AI solution is used within a broader system. These issues were handled by offering inclusive definitions of AI and ADM to participants and asking questions relating to both their use and awareness of their use:

> On behalf of the Equality Ombudsman, the use of artificial intelligence and other automated decision-making in recruitment processes in working life is investigated here. Automatic decision-making encompasses algorithmic and automated decision-making processes both with and without artificial intelligence. In the term, we include both fully automated decision-making and automated processes used as decision support. (translated from Swedish)

Similar definitional issues can be identified concerning ADM. In studies of public administration, the term "automated decision-making" is often used as a holistic concept to refer to various forms of automation in decision support (Roehl, 2022). Simpler rule-driven algorithmic systems for decision-making can thus be included; this means there is a longer history of the use of automated decisions, which may have been made daily but can still lead to challenges, as in the case of municipal income support (Kaun, 2022). However, the difference between automation for *decision support*, where a human has the final say, and fully automated *decision-making* (with no human intervention) can be crucial, not least because studies on the type of automation in the public sector that people generally trust show that support is far greater for ADM as a non-fully autonomous decision-support mechanism, where a human makes the final decision, than as a fully automated practice (Insight Intelligence, 2022). This type of literature often points to the significance and consequences of the development towards the quantification, or so-called datafication, of a range of processes, such as in welfare (Kaun et al., 2023) or in quantifying work and workers (Ajunwa, 2023).

2.2. The Risks of Discrimination of AI and ADM

AI and ADM are sometimes promoted as an unbiased alternative to human judgement (analysed in terms of "regimes of justification" by Dencik & Stevens, 2021). The argument goes that because machines do not harbour explicit or subconscious preferences for individual traits or characteristics, or because they are not prejudiced against groups of people, they cannot unfairly discriminate between them. Related to this is the assertion that because machines do not grow tired or dissatisfied with their work, they are unable to make the same mistakes or errors in judgement that humans often make.

There is now a significant body of work that shows the limits of this argument (e.g., Benjamin, 2019; Crawford, 2021; Noble, 2018). Because AI and ADM are developed by humans and trained on data that reflect human language and society, they internalise human biases and inequalities and come to reflect pre-existing problems in prevailing social structures (D'Ignazio & Klein, 2023; Larsson, 2019; Larsson et al., 2023b). Complex AI systems, like the recommender systems in search engines, can encode biases that are deeply embedded in historic and persistent distributions of opportunity. When internet studies scholar Noble (2018) searched for

"black girls," she was astonished to discover that the first result returned by Google was for a pornography site. Her subsequent research into search results for websites, images, and geographical locations revealed systematic lack of credible and reliable information about women and people of colour, reinforcing racist and oppressive stereotypes, and the structural relations of domination that underlie them. Only when her work became a controversy for Google did they do anything to amend their algorithms and search results.

But it is not only complexity that is at issue here. Large language models trained on massive volumes of unstructured data enclose statistical relationships between words and groups of words within neural networks that have been argued to be inherently opaque (Peters, 2023). Just as the training data for these systems is rife with human prejudice, so do the systems themselves become prejudiced. When researchers asked such a system to complete the sentence "Man is to computer programmer as woman is to _____" its response was "homemaker" (see Bolukbasi et al., 2016; Caliskan et al., 2017). These kinds of problematic associations are the very same issues that beset OpenAI's ChatGPT, Google's Bard, and Microsoft's Bing (cf. Ferrara, 2023). They cannot easily be removed from trained models, and so developers have taken to placing evolving controls on model outputs. Even if prejudiced statements are fully and effectively removed from the training data, the output of large language models will remain inconsistent and unpredictable, and so always at risk of transgressing widely held values (Bender et al., 2021).

These kinds of problems also affect computer vision systems. The AI models at the core of these systems are trained on datasets of labelled images or videos. These data have often been produced by thousands of individuals who bid on small, on-demand jobs posted to crowdsource work marketplaces such as Amazon's Mechanical Turk—part of the gig economy touched on in the introduction section. While this allows any single data point to be cross-checked by several persons, audits of datasets that have been produced in this manner still reveal problems and inconsistencies (Crawford, 2021). These occur due to commonly repeated errors and biases in the tagging, and due to the unequal distribution of images within the dataset. In their study of facial recognition systems, Buolamwini and Gebru (2018) found that the lack of representation of female and dark-skinned faces within training data meant that the resultant models performed more poorly on women and people of colour. These results were compounded by intersecting absences and Black women were the most poorly recognised of all.

It is, however, far from clear that such problems with AI-driven image detection can be solved by simply fixing the data. In 2015, Google's photo web app automatically added "gorilla" tags to the faces of an African American couple. This was an unintended and embarrassing error that Google was quick to respond to by removing the tag (Russell, 2019). The problem is not only that the AI performed poorly on black faces or that the choice of the tag was racist to begin with; the problem is also that the meaning of this error can only be understood in terms of the history of American slavery and racial oppression, the justification for which was through scientific racism and the dehumanisation of Black people (Benjamin, 2019). This is not something that the model can know. The context of the error, whatever its cause, escapes the machine entirely.

The causes and likelihood of algorithmic bias and discrimination within complex technical systems are often hard to pin down. This is because such systems are not only complex, they also include trade secrets, and interests which do not wish to be (or should not be) openly auditable. Or, as Morondo Taramundi (2022, p. 73), articulates it: "Technical complexity, together with proprietary interest and economic calculations, makes it very difficult to understand exactly how algorithms discriminate." This is a particularly relevant concern for

AI and ADM used in hiring related to the large-scale tech platforms that many depend on in their indirect use of these technologies. Issues related to transparency and accountability are recognised challenges for AI systems and ADM as such (Larsson & Heintz, 2020), which can pose problems both for the governance of AI systems (e.g., Novelli et al., 2023; Taeihagh, 2021), as well as for large digital platforms (Geng, 2023; Kim & Moon, 2021). We return to this in Section 5. This research examines this accountability challenge linked to proprietary platforms and the associated knowledge gap when it comes to the use of AI and ADM as tools, methods, and services.

3. Methods and Data

The research was formulated by two of the authors in collaboration with the research company Novus and the Swedish Equality Ombudsman. In preparation for the questionnaire phase of the study, six unstructured scoping interviews were conducted with area experts. These helped clarify the questions to be included in the telephone questionnaire aimed at the two target groups: recruitment agencies and companies with more than 100 employees. The questionnaires were conducted via telephone calls to access an occupied group of professionals who are unlikely to take the time to answer text-based questionnaires sent to them via email (or mail). While the questionnaire responses entail self-reported data, we have no reason to believe that the respondents were biased or misleading in their answers.

The number of contact attempts per person for both groups was eight to ten. This strategy was motivated by the fact that this type of respondent is relatively unavailable, the population relatively small, and it was extremely important to get hold of the right respondent. The interview time for recruitment companies was on average 10 minutes and 12 minutes on average for the major employers.

3.1. Target Group 1: Recruitment Agencies

For recruitment agencies, we approached 50 of Sweden's largest recruitment agencies. Many recruitment agencies seem to have a policy of not responding to requests for information on how the services they offer are carried out. Of the 50 recruitment agencies contacted, 10 agreed to answer the questionnaire. Of these 10, nine recruit more than 50 people per year, and one recruits 11–20 people per year.

To ensure respondents had sufficient knowledge to answer the questions, the interviews were primarily conducted with chief technology officers or an equivalent HR manager. We started from a list provided by employer organisation Kompetensföretagen (part of Almega) of the 50 largest employment agencies (which includes recruitment agencies) in Sweden based on turnover. This list is published every quarter and we used the most recent edition.

3.2. Target Group 2: Major Employers

For major employers, 100 major Swedish employers responded to our questionnaire. This sample was comprised of 50 representatives of companies with 100–249 employees and 50 representatives of employers with more than 250 employees, of which 23 have more than 500 employees (see Table 1).

Table 1. Distribution of sample of major employers.

Number of employees in each participating company	Number of questionnaires answered
100–149 employees	22
150–249 employees	28
250–499 employees	27
500+ employees	23
Total	100

Statistics Sweden estimates there to be 1,904 companies in the range of 100–249 employees and 1,205 in the range of 250 employees and above (Statistics Sweden, 2023). Of the companies within the sample, about half (49%) recruit more than 50 new employees per year. Given the population size, a sample of 100 respondents gives more than 95% confidence in a 10% margin of error in the responses. It was determined that this would be sufficient for the purpose of this research, which was not to establish strong statistical relationships but to map potential discrimination risks. The data did not include any personal identifiers or personal information, only reported use by a particular organisation. This study was designed in collaboration with the public authority tasked with battling unlawful discrimination, the Swedish Equality Ombudsman, in a way that it should adhere to principles of ethical conduct in research as they are regulated in Sweden: In Sweden, existing law on ethical conduct in research, specifically the Ethics Review Act (but often in combination with the GDPR), aims to protect individual research subjects and respect for human dignity. The law defines what kinds of research must be approved by the Ethics Review Authority (Etikprövningsmyndigheten) before being carried out (Articles 3 and 4). The present study clearly does not meet this threshold and therefore did not require approval from the Ethics Review Authority. Furthermore, the Ethics Review Authority has recently issued guidance on what risks to consider before conducting research (Görman & Etikprövningsmyndigheten, 2023). Of particular relevance for this study is the information given to the respondents and the necessity to protect the integrity of the respondent through the anonymisation of personal data. In keeping with the Ethics Review Act and this supplementary guidance, independent ethical approval was not needed to conduct this study.

The questionnaire was geared towards HR managers or other senior managers with clear recruitment responsibilities or equivalent positions. Researchers took care to ensure that they were speaking to the person responsible and were often transferred several times within a company until they were put in contact with someone suitable. While these representatives may not accurately depict overall sentiment within the company, they are expected to have a greater level of awareness of recruitment practices and technologies than the company average. As these respondents act as a proxy for their company, care is taken to limit the analysis and not overstate the conclusions.

3.3. The Questionnaires

The two respondent groups received similar questionnaires. For the recruitment agencies, there were eight questions, with five follow-up questions depending on what the respondents answered. For the major employers, there were nine questions with four follow-up questions. The most important questions for this article can be found in Figures 1 through 4.

4. Results and Analysis

The analysis combines the descriptive results from the surveys with the theoretical needs required to analyse the importance of AI and automation concepts and how risks of discrimination can interact with technical solutions and in relation to already established knowledge in the field. The analysis thus includes the differences between perceived and reported use, direct and indirect use, and the risks associated with a lack of transparency in both AI systems and the services of large-scale platforms, along with questions about the allocation of accountability for any resulting bias or discrimination.

4.1. Do the Respondents Really Not Use AI?

Recruitment agencies were asked about the presence of these technologies in services offered to clients. Levels of perceived use of ADM within recruitment management systems were low, with eight out of ten responding that usage was to a fairly low extent, a very low extent, or not at all. Similar trends applied to the use of AI, where eight out of ten respondents said that use was very low or not at all. Among the major employers, 87% responded that the use of ADM/decision support in recruitment was fairly low or very low (Figure 1).

Given that respondents are expected to have slightly different perceptions of what constitutes AI and ADM, we cannot assume that this is an accurate representation of the actual use of these technologies. Rather, the results offer an overview of their *perceived* use amongst company representatives, who are assumed to have higher-than-average awareness levels. When we asked about specific applications and services, reported use was much higher.

4.2. Direct and Indirect Use of AI and ADM

Further questions in both the recruitment agency questionnaire (Figure 2) and in the major employer questionnaire were designed to better measure the actual use of AI and ADM (see Figure 3), and ultimately also revealed a distinction between direct and indirect use of AI and ADM. When recruiters were asked if

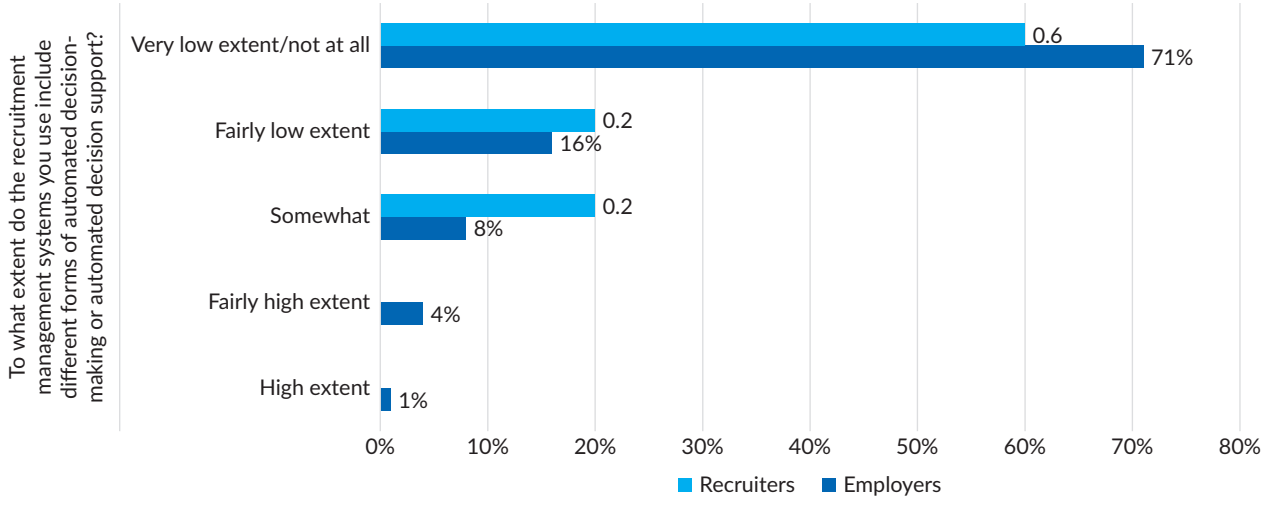

Figure 1. Do recruitment agencies and major employers use AI in their recruitment?

they used AI or another ADM to automatically match job seekers with jobs, seven out of ten of the respondents confirmed that they did (Figure 2). Since this is a clear reference to the variants of social media that, according to several studies, are common in job-matching contexts, such as LinkedIn, we can conclude that the respondents use AI functionalities with a large element of automation in their recruitment process, although they do not see it as such—hence the lower perceived use. Eight out of ten also said that they received information about suitable candidates from social media profiles, such as Facebook, LinkedIn, Instagram, TikTok, Twitter, and others (see Figure 2), which also make use of AI in their search and recommender operations. Most recruitment agencies (9 out of 10) tested candidates through games, IQ tests, or personality tests.

All, however, indicated that they used existing job platforms, such as the Swedish Public Employment Service or Monster. These job platforms offer search functions that can rank hits against either job postings (for job applicants) or jobseekers' résumés (for recruiters), using varying degrees of automation. So, although the agencies do not explicitly report this in the question about perceived AI use, all of them do, in fact, use some kind of ADM in their recruitment work, often through a third-party service.

Similarly, respondents in the large company questionnaire reported extensive use of digital platforms in their hiring practices, including recruitment platforms (80%) such as ReachMee, Varbi, LinkedIn Recruiter, and Teamtailor, and social media (55%), e.g., Facebook, LinkedIn, Instagram, and similar (see Figure 3).

Outsourcing of recruitment is also very common. 69% of the Swedish employers stated that they outsource some of their recruitment processes to a recruitment agency (see Figure 3). While the vast majority report that they use a recruitment platform and about half that they use social media in their recruitment processes, it is possible that some respondents interpreted outsourcing as one of these services. The extent to which and in what combination the respondents use these different services is not clear from the data.

The analytical challenge, which is related to the ambiguity involved in defining both AI and ADM, is visible in these questionnaire results. While it appears as though ADM is frequently used among recruiters (even if this is not fully acknowledged), it is unclear whether this is true of major employers. Recruitment management

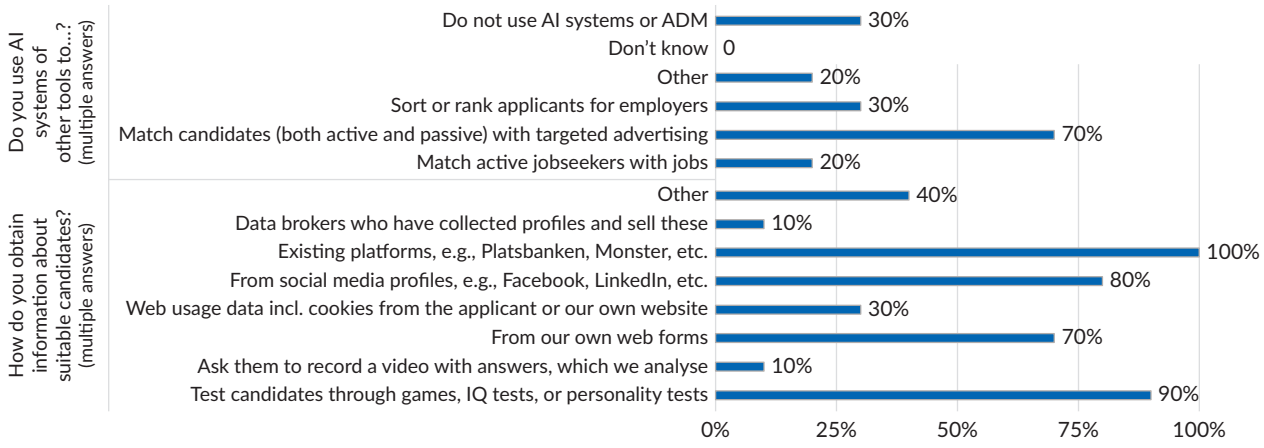

Figure 2. What do recruitment agencies use AI systems for, and how do they find information about candidates?

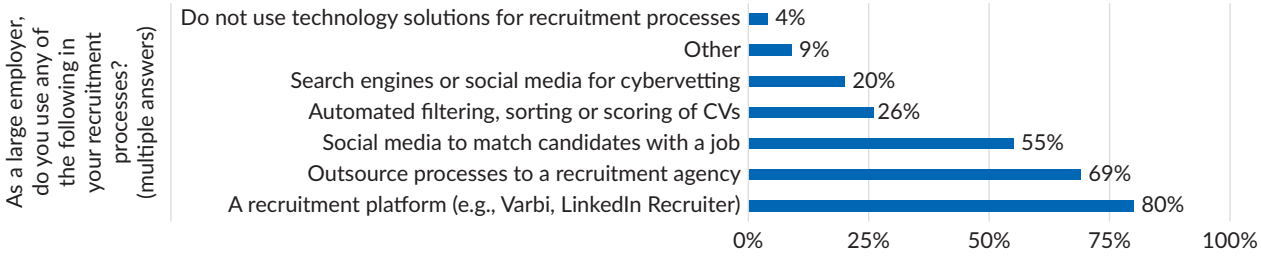

Figure 3. Which tools do major employers use in their recruitment processes?

platforms, such as ReachMee and Varbi, are integrated solutions developed to streamline job posting and candidate management. They are not marketed as AI-powered solutions but use forms of automated analysis for decision support. They can include forms of ADM, as we have defined the term. LinkedIn Recruiter, for instance, has publicised its use of AI to match candidates to jobs and to help identify potential recruitment opportunities (LinkedIn, 2018).

4.3. Accountability

Here, we use transparency to refer to auditability, access to databases, end-users' information skills, or the need for documentation about the algorithms used, and the goals governing the AI model (e.g., Larsson & Heintz, 2020; Larsson et al., 2023a). When the recruitment agencies were asked whether they told candidates that they used AI or ADM tools, six declined to answer, one said that they did not declare this but probably should, while three said that they did not know. This is clearly a challenge—compounded by both the ubiquity and extensive indirect use of tools with varying degrees of automation.

When the major employers were asked who they believe is responsible for discrimination occurring because of the use of an AI system or ADM, 84% stated that the responsibility lies with the company using the system (see Figure 4). At the same time, there is a clear and immediate risk that neither companies nor recruiters have any means of checking how the AI systems that are indirectly used ensure that they do not discriminate when matching candidates via social media or third-party recruitment services. The issue is complicated both by the inherent challenges in auditing AI systems and by the low transparency of platform companies (although the latter may change under the DSA, see Söderlund et al., 2024).

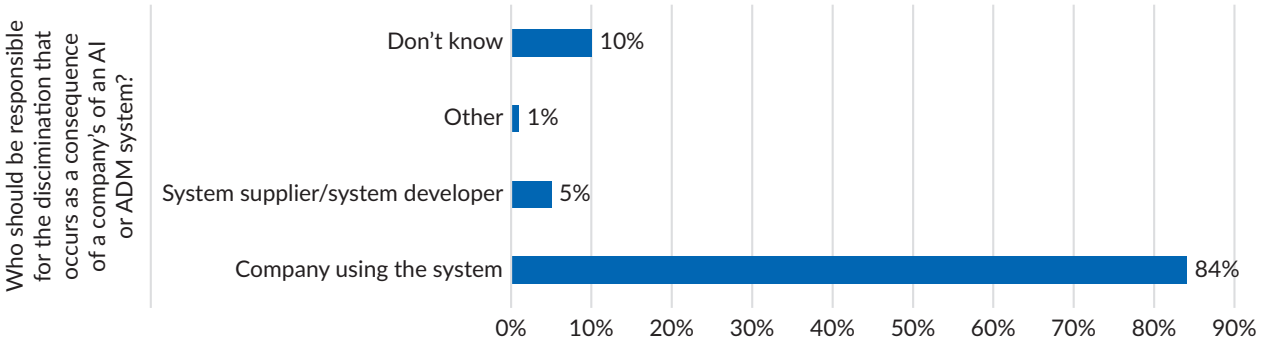

Figure 4. Who is accountable?

Overall, this data points to the extensive use of AI and ADM tools for recruitment both among recruitment agencies and among major employers in Sweden. It is interesting to note that recruiters (8 out of 10) and major employers (77%) say that their use of AI and ADM is low or very low, but that they make extensive use of tools and platforms like LinkedIn Recruiter and similar that explicitly market themselves as AI-enabled. Transparency and responsibility remain largely hypothetical concerns: Although major employers believe that they are themselves responsible for any discrimination, both extensive indirect use and laws like the DSA and the forthcoming AI Act suggest that the responsibility of platforms providing automated tools is not to be ignored.

5. Discussion

Taken together, these results indicate that there is low awareness of the everyday use of AI and ADM in the Swedish workplace by recruiters. This may indicate a lack of attention to the role played by these technologies in tools used every day. Every time a web search is performed (e.g., through Google), an automated recommendation system is used. Every time a user scrolls through someone's social media feed (e.g., on Facebook), an AI system determines what they see. Every ad that is seen on these websites and digital platforms has been selected by a complex automated model.

In this study, the difference between perceived and reported use of AI is striking and perhaps a testament to the ubiquity of AI tools in the workplace today. This study supports prior work that suggests that the increased prevalence of AI and other ADM in the workplace is not likely to lead to increased awareness. Instead, ubiquity leads to technologies being taken for granted and receding into the background (cf. Kaun, 2022; Lomborg, 2022). This growing body of evidence suggests there is a risk that these technologies will integrate into everyday life and not only become unreflexively trusted by the people who use them, but even ignored.

5.1. Indirect Use of AI and ADM

A further case in point lies in the distinction we made above between direct and indirect use of AI and ADM. Here, we understand direct use to occur when solutions have been significantly developed by the companies that use them, and indirect use to occur when a technical solution like AI is packaged into a service used by a company but provided by a third party, e.g., LinkedIn.

Indirect use is most apparent when one considers the use of technology solutions in recruitment processes. For instance, the finding that 26% of major employers said they used automated résumé filtering, sorting, or scoring (see Figure 3). Although this option was intended to measure direct use, it is more likely to be a mixture of both direct and indirect use. Much more common is that companies use a recruitment platform (with 80% of respondents saying that they do; see Figure 3), into which this type of candidate sorting can be integrated. 55% of respondents reported using social media (including LinkedIn) to match candidates with jobs, and 20% reported using search engines or social media to find background information about potential candidates (see Figure 3). These results indicate that indirect use of AI and other ADM is significantly higher than direct use and that these uses are usually tied to a software solution for which some support is available.

There are two further analytical distinctions to be made here. The first is that there is some overlap between the categories of direct and indirect use and that many of the options we offered the survey respondents are not unambiguously one or the other. Without explicitly asking whether the systems in question were developed in-house, the best we can offer is an estimate of the extent to which they fall in one category or the other. The second challenge is that indirect use covers a wide variety of products and services. A further distinction can be made between those uses that are easily accessible to the public (e.g., Google search) and those that were acquired by the company and therefore can be expected to come with support (e.g., Varbi). Another can be made between products that include AI as an additional feature (e.g., Microsoft Word) and those with a feature set largely built around AI models (e.g., LinkedIn Recruiter). The latter is associated with the potential risks of discrimination stemming from how AI models are trained, but with the addition that it is difficult for any external actor to audit proprietary and large-scale systems. Most indirect uses fall somewhere between these extremes.

In the findings detailed above, respondents were quick to highlight when they made direct use of AI or ADM but largely ignored indirect use. Given the format of the questionnaires, it is unclear whether this was out of ignorance or denial, but it is a question deserving of further research.

5.2. The Ubiquity of LinkedIn

LinkedIn, given its ubiquity in recruitment processes (Ajunwa & Greene, 2019; Laukkarinen, 2023; Ruparel et al., 2020), is of particular interest here. As it is not possible to completely differentiate LinkedIn from other recruitment platforms in the questionnaire results (especially as many firms use a mix of recruitment methods), the extent to which AI-powered solutions are used in recruitment processes cannot be measured definitively. Nevertheless, the results indicate that the use of LinkedIn is common, and one interpretation of the results is that around two-thirds of companies that use a recruitment platform use that one in particular. This means that companies' assessment of discrimination risks through the use of AI or other ADM relies heavily on their belief that LinkedIn manages these risks. This suggests that more attention should be given to the role of large-scale digital platforms in addressing discrimination issues in general, and to the measures they take to reduce the risks of discrimination. Although the DSA requires significant transparency from services like LinkedIn, how and if this transparency affects firms that use their services is deserving of further empirical attention.

5.3. Lack of Transparency and (Distributed) Accountability

These concerns around indirect use lead also to a renewed emphasis on the importance of transparency during these processes. Issues related to transparency and accountability are recognised problems, both as governance challenges for AI systems and large digital platforms (e.g., Geng, 2023; Novelli et al., 2023). These are particularly salient here given that the questionnaire data indicate that AI and ADM are used both directly and indirectly, with implications for transparency and accountability. This indirect use may also lead to lower awareness of the risks of discrimination in the use of AI and other ADM in the workplace (e.g., Khatry, 2020). This poses a particular challenge in terms of oversight, where laws in the EU (specifically the DSA) understand large platforms as increasingly responsible for biased content, but not necessarily for the consequences of this content.

Given that both recruitment agencies and major employers underestimate the extent to which they use AI and ADM services while making extensive use of services that have these technologies built in, greater awareness in general is called for. However, opacity in algorithms, and especially complex self-learning AI systems, makes it hard (if not impossible) to assess how and to whom accountability should be allocated (cf. Novelli et al., 2023). While respondents to our questionnaire overwhelmingly thought that the company using the system should be held responsible for bias (84%; see Figure 4), this risks allocating responsibility to actors who do not influence the underlying algorithmic models. This is possible both in direct use where a service has not been developed in-house, and a given when talking about indirect use of services offered by large digital platforms. More transparency would help in offering a more nuanced approach to accountability in what we think of as being distributed rather than shared (Dignum, 2019).

Regulations that have attempted to increase transparency (including the DSA) have faced numerous stumbling blocks. Among them is an inherent conflict of interest between the need for transparency for a supervisory authority—for example, to be able to review how an AI system is developed—and companies' interest in not disclosing how their products work in a competitive market (Larsson & Heintz, 2020). Still, another is the argument that complex algorithms, for instance, large language models (LLMs), are inherently opaque and that it is just not technically feasible to be fully transparent (cf. Peters, 2023). Transparency brings with it the risk of abuse and workarounds—where users learn to subvert responsibility (cf. De Laat, 2018). While this article has focussed on recruitment firms and major employers, transparency alone does not give small employers the resources to either comply with or remedy concerns around bias and risk in AI and ADM use.

Large digital platforms, of which LinkedIn is one, highlight that transparency is important not just at the level of how data are handled and how algorithms draw the conclusions that they do (e.g., Larsson et al., 2021). Rather, the extensive use of such platforms (including unreflexively) adds to a shift in the balance of power in a society, where societal functions are controlled by a few globally active—largely American-anchored—technology conglomerates (Larsson, 2021; van Dijck et al., 2018).

6. Conclusion

This study has aimed to contribute to knowledge about (a) the extent to which AI and ADM are used in recruitment processes, (b) usage perception, and (c) questions around accountability in light of concerns around AI and ADM leading to bias in recruitment. While recruitment companies and major employers overwhelmingly said that their use of ADM/decision support in recruitment was fairly low or very low, in practice all respondents of both categories made use of third-party digital platforms, including social media and recruitment services like LinkedIn, that use AI or ADM to varying degrees. While it is hard to measure, then, the exact extent to which recruitment agencies and major employers use AI or ADM in recruitment (given overlaps and the variety of definitions and degrees of AI use) differs depending on whether one talks about direct use or indirect use.

This extensive indirect use of AI and ADM raises concerns about awareness and reflexivity during recruitment processes. It also has implications for accountability and oversight of the discrimination and bias risks implicit in AI and ADM use. While there have been significant advances made in compelling transparency among, for instance, large digital platforms (through the DSA) and high-risk uses of AI (through the AI Act), this does not necessarily solve the question of who is accountable for bias or discrimination that may result from AI or ADM

use in recruitment. While large employers were quick to say that they should themselves be held accountable (84%), transparency would make it easier to decide how, and to whom, accountability and responsibility should be allocated.

Overall, the use of AI and ADM is already well-established in Sweden, and its prevalence is likely to be an exemplar case of what recruitment looks like in the rest of the world. While the exact numbers and platform names might differ, the issues raised here around awareness, indirect use, transparency, and accountability are likely to be generalizable to other contexts. The use of questionnaires here to access busy professionals allowed us to gain a snapshot view of the issues. The exact dividing lines between the (direct and indirect) use of technologies with different features, as well as the motivations and trade-offs that inform AI and ADM use, are beyond the scope of this data. However, the results here suggest that these more granular issues around AI and ADM use in recruitment are deserving of further attention by both researchers and policymakers.

Acknowledgments

The authors would like to extend their gratitude to the Swedish Equality Ombudsman (DO) as well as the Research Institute for Sustainable AI (RISAI) for significant assistance in the execution of this study.

Funding

This study has in part been funded by (a) the Swedish Equality Ombudsman and (b) the Wallenberg AI—Autonomous Systems and Software Program–Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The data can be made available to bone fide researchers upon request.

References

Ajunwa, I. (2023). *The quantified worker: Law and technology in the modern workplace*. Cambridge University Press.

Ajunwa, I., & Greene, D. (2019). Platforms at work: Automated hiring platforms and other new intermediaries in the organization of work. In S. P. Vallas & A. Kovalainen (Eds.), *Work and labor in the digital age* (Vol. 33, pp. 61–91). Emerald Publishing.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim code*. Polity Press.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS'16: Proceedings of the 30th international conference on neural information processing systems* (pp. 4349–4357). Curran Associates Inc.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender

classification. In X. Alameda-Pineda, M. Redi, E. Celis, N. Sebe, S.-F. Chang (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91). ACM.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

D'Ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT Press.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, *31*(4), 525–541.

Dencik, L., & Stevens, S. (2021). Regimes of justification in the datafied workplace: The case of hiring. *New Media & Society*, *25*(12), 3657–3675. https://doi.org/10.1177/14614448211052893

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (the AI Act) and amending certain Union Legislative Acts* (COM/2021/206 final).

European Commission. (2022). *Digital Services Act* (Regulation (EU) 2022/2065, DSA).

European Commission. (2023). *Artificial intelligence—Questions and answers*.

Ferrara, E. (2023). *Should ChatGPT be biased? Challenges and risks of bias in large language models*. Arxiv. https://arxiv.org/abs/2304.03738

Gasser, U., & Almeida, V. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58–62.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128.

Geng, Y. (2023). Transparency for what purpose? Designing outcomes-focused transparency tactics for digital platforms. *Policy & Internet*. Advance online publication. https://doi.org/10.1002/poi3.362

Görman, U., & Etikprövningsmyndigheten. (2023). *Vägledning—Om etikprövning av forskning på människor*. Etikprövningsmyndigheten.

Hagendorff, T., & Wezel, K. (2020). 15 challenges for AI: Or what AI (currently) can't do. *AI & Society*, *35*, 355–365.

Ilsøe, A., & Söderqvist, C. F. (2022). Will there be a Nordic model in the platform economy? Evasive and integrative platform strategies in Denmark and Sweden. *Regulation & Governance*, *17*(3), 608–626.

Insight Intelligence. (2022). *The Swedish people and AI: Swedish people's attitudes towards artificial intelligence 2022*.

Jones, P. (2021). *Work without the worker: Labour in the age of platform capitalism*. Verso.

Kaun, A. (2022). Suing the algorithm: The mundanization of automated decision-making in public services through litigation. *Information, Communication & Society*, *25*(14), 2046–2062.

Kaun, A., Lomborg, S., Pentzold, C., Allhutter, D., & Sztandar-Sztanderska, K. (2023). Crosscurrents: Welfare. *Media, Culture & Society*, *45*(4), 877–883. https://doi.org/10.1177/01634437231154777

Kelly-Lyth, A. (2021). Challenging biased hiring algorithms. *Oxford Journal of Legal Studies*, *41*(4), 899–928.

Khatry, S. (2020). Facebook and Pandora's box: How using big data and artificial intelligence in advertising resulted in housing discrimination. *Applied Marketing Analytics*, *6*(1), 37–45.

Kim, K., & Moon, S. I. (2021). When algorithmic transparency failed: Controversies over algorithm-driven content curation in the South Korean digital environment. *American Behavioral Scientist*, *65*(6), 847–862.

Komljenovic, J. (2019). Linkedin, platforming labour, and the new employability mandate for universities. *Globalisation, Societies and Education*, *17*(1), 28–43.

Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et Société*, *103*(3), 573–593.

Larsson, S. (2021). Putting trust into antitrust? Competition policy and data-driven platforms. *European Journal of Communication*, *36*(4), 391–403.

Larsson, S., Haresamudram, K., Högberg, C., Lao, Y., Nyström, A., Söderlund, K., & Heintz, F. (2023a). Four facets of AI transparency. In S. Lindgren (Ed.), *Handbook of critical studies in artificial intelligence* (pp. 445–455). Edward Elgar Publishing.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Larsson, S., Jensen-Urstad, A., & Heintz, F. (2021). Notified but unaware: Third-party tracking online. *Critical Analysis of Law*, *8*(1), 101–120.

Larsson, S., Liinason, M., Tanqueray, L., & Castellano, G. (2023b). Towards a socio-legal robotics: A theoretical framework on norms and adaptive technologies. *International Journal of Social Robotics*, *15*, 1755–1768. https://doi.org/10.1007/s12369-023-01042-9

Laukkarinen, M. (2023). Social media as a place to see and be seen: Exploring factors affecting job attainment via social media. *The Information Society*, *39*(4), 199–212.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering* (pp. 14–16). ACM.

LinkedIn. (2018). *An introduction to AI at LinkedIn*. https://engineering.linkedin.com/blog/2018/10/an-introduction-to-ai-at-linkedin

Lomborg, S. (2022). Everyday AI at work: Self-tracking and automated communication for smart work. In S. Pink, M. Berg, D. Lupton, & M. Ruckenstein (Eds.), *Everyday automation: Experiencing and anticipating emerging technologies* (pp. 126–139). Routledge.

Morondo Taramundi, D. (2022). Discrimination by machine-based decisions: Inputs and limits of anti-discrimination law. In B. Custers & E. Fosch-Villaronga (Eds.), *Law and artificial intelligence: Regulating AI and applying AI in legal practice* (pp. 73–85). TMC Asser Press.

Myers West, S., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race, and power in AI*. AI Now Institute.

Newlands, G. (2022). Anthropotropism: Searching for recognition in the Scandinavian gig economy. *Sociology*, *56*(5), 821–838. https://doi.org/10.1177/00380385211063362

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & Society*. https://doi.org/10.1007/s00146-023-01635-y

Pasquale, F. (2020). *New laws of robotics: defending human expertise in the age of AI*. The Belknap Press of Harvard University Press.

Peters, U. (2023). Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. *AI and Ethics*, *3*(3), 963–974.

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Roehl, U. B. U. (2022). Understanding automated decision-making in the public sector: A classification of automated, administrative decision-making. In G. Juell-Skielse, I. Lindgren, & M. Åkesson (Eds.), *Service automation in the public sector. Concepts, empirical examples and challenges* (pp. 35–63). Springer Nature.

Ruparel, N., Dhir, A., Tandon, A., Kaur, P., & Islam, J. U. (2020). The influence of online professional social media

in human resource management: A systematic literature review. *Technology in Society*, *63*. https://doi.org/10.1016/j.techsoc.2020.101335

Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Books.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 458–468). ACM.

Statistics Sweden. (2023). *Antalet anställda i snabbväxande storföretag kommer att öka*. https://www.scb.se/hitta-statistik/temaomraden/sveriges-ekonomi/fordjupningsartiklar_Sveriges_ekonomi/antalet-anstallda-i-snabbvaxande-storforetag-kommer-att-oka

Söderlund, K. & Engström, E. & Haresamudram, K. & Larsson, S., & Strimling, P. (2024). Regulating high-reach AI: On transparency directions in the Digital Services Act. *Internet Policy Review*, *13*(1). https://doi.org/10.14763/2024.1.1746

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157.

Thylstrup, N. B. (2018). *The politics of mass digitization*. MIT Press.

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197–208.

van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*. https://doi.org/10.1016/j.clsr.2021.105567

Woodcock, J., & Graham, M. (2020). *The gig economy: A critical introduction*. Polity Press.

Zhang, L., & Yencha, C. (2022). Examining perceptions towards hiring algorithms. *Technology in Society*, *68*. https://doi.org/10.1016/j.techsoc.2021.101848

About the Authors

Stefan Larsson is a senior lecturer and associate professor in technology and social change at Lund University, Sweden, Department of Technology and Society. He is a lawyer and socio-legal researcher who holds a PhD in sociology of law as well as a PhD in spatial planning. He leads a multidisciplinary research group that focuses on issues of trust and transparency and the socio-legal impact of autonomous and AI-driven technologies in various domains, such as consumer markets, the public sector, and in health and social robotics.

James Merricks White is a postdoctoral researcher in the Department of Technology and Society, Lund University, where he is studying trust and trustworthiness in AI consumer markets. His research interests include expertise, knowledge/ignorance, and standardisation, with a focus on the International Organization for Standardization (ISO) and the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). He has recently been awarded a VR grant (2023-01549) to study risk governance mechanisms in the AI Act.

Claire Ingram Bogusz is an associate professor (docent) in the Department of Informatics and Media, with a background in Law, philosophy, and economics. She conducts research into how code-based technologies (e.g., code itself, digital platforms and infrastructures, learning machines, and digital automation) change how we work, how organising occurs, and what these things mean for us as individuals and as members of increasingly polarised and unequal societies.

Social Inclusion
2024 • Volume 12 • Article 7543
https://doi.org/10.17645/si.v12.7543

cogitatio

ARTICLE

Open Access Journal

Intersectionality in Artificial Intelligence: Framing Concerns and Recommendations for Action

Inga Ulnicane

University of Birmingham, UK

Correspondence: Inga Ulnicane (i.ulnicane@bham.ac.uk)

Submitted: 18 August 2023 **Accepted:** 26 February 2024 **Published:** 18 April 2024

Issue: This article is part of the issue "Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination" edited by Derya Ozkul (University of Warwick), fully open access at https://doi.org/10.17645/si.i236

Abstract

While artificial intelligence (AI) is often presented as a neutral tool, growing evidence suggests that it exacerbates gender, racial, and other biases leading to discrimination and marginalization. This study analyzes the emerging agenda on intersectionality in AI. It examines four high-profile reports dedicated to this topic to interrogate how they frame problems and outline recommendations to address inequalities. These four reports play an important role in putting problematic intersectionality issues on the political agenda of AI, which is typically dominated by questions about AI's potential social and economic benefits. The documents highlight the systemic nature of problems that operate like a negative feedback loop or vicious cycle with the diversity crisis in the AI workforce leading to the development of biased AI tools when a largely homogenous group of white male developers and tech founders build their own biases into AI systems. Typical examples include gender and racial biases embedded into voice assistants, humanoid robots, and hiring tools. The reports frame the diversity situation in AI as alarming, highlight that previous diversity initiatives have not worked, emphasize urgency, and call for a holistic approach that focuses not just on numbers but rather on culture, power, and opportunities to exert influence. While dedicated reports on intersectionality in AI provide a lot of depth, detail, and nuance on the topic, in the patriarchal system they are in danger of being pigeonholed as issues of relevance mainly for women and minorities rather than part of the core agenda.

Keywords

artificial intelligence; data; diversity; feminism; framing; gender; intersectionality; politics; power; technology

1. Introduction

Amid the hype and excitement surrounding the potential economic and social benefits of artificial intelligence (AI), uncomfortable questions regarding the wide-ranging problematic impacts of AI are multiplying (Radu, 2021; Schiff, 2023; Taeihagh, 2021; Ulnicane et al., 2021). While AI is often perceived as neutral and objective, growing evidence suggests that it reinforces and exacerbates human biases and stereotypes leading to disadvantages and discrimination based on gender, race, ethnicity, age, and other characteristics (Allhutter et al., 2020; Benjamin, 2019; Broussard, 2023; Browne et al., 2023; D'Ignazio & Klein, 2020; Noble, 2018; Søraa, 2023; Ulnicane & Aden, 2023). Numerous problematic examples include the hiring tool that discriminates against women, racial discrimination being built into sentencing algorithms, targeted ads, chatbots, voice assistants, and robots replicating gender and racial biases, or facial recognition working poorly on black female faces (Angwin et al., 2016; Buolamwini & Gebru, 2018; Collett & Dillon, 2019; Young et al., 2021).

These examples are closely related to the diversity crisis in AI, where founders as well as employees of major companies largely come from a homogenous group of white men from well-off socio-economic backgrounds and build AI systems according to their views and stereotypes (Little & Winch, 2021; S. M. West et al., 2019; Young et al., 2023). Moreover, this is happening at a time when illiberal pushback against gender equality in technology is encountered at the highest levels of political decision-making as well as through widely publicized initiatives from individuals at big tech companies (Schopmans & Cupac, 2021; S. M. West et al., 2019). Concerns have been raised that discriminatory AI systems might offset earlier advances made towards gender equality (UNESCO, 2020). These examples and concerns highlight problematic political and social impacts of AI, its reinforcement of existing power relationships, structural inequalities, and the highly unequal distribution of rewards and costs of new technologies across various social groups.

Against this background, this article examines an emerging agenda to address inequality issues in AI. To do that, it uses the concept of intersectionality, which deals with multiple inequalities arising from the interaction of various social identities, such as gender, race, ethnicity, and socio-economic background, contributing to the marginalization and disadvantaging of less powerful groups (Crenshaw, 1991; Verloo, 2006). In particular, this study asks how the problems of intersectionality in AI are framed and what recommendations have been outlined to address them. To get a good overview of the problems and solutions, four dedicated reports on intersectionality in AI are analyzed: Collett and Dillon (2019), UNESCO (2020), S. M. West et al. (2019), and Young et al. (2021). This article on intersectionality and AI contributes to other studies on gender and bias in AI policy (Guevara-Gómez et al., 2021; Rönnblom et al., 2023; Ulnicane & Aden, 2023), unpacking how inequalities and discrimination are understood and tackled in documents prepared by various experts and stakeholders aiming to influence and set AI policy.

This article proceeds as follows. First, it outlines the conceptual and methodological framework introducing the key concepts of AI and intersectionality (Section 2.1), as well as explaining the choice of empirical material (Section 2.2). Second, it examines the framing of concerns and recommendations in the four selected reports (Section 3). Finally, the findings from the analysis are discussed (Section 4).

2. Conceptual and Methodological Framework

2.1. Key Concepts: AI and Intersectionality

AI is a highly contested and political concept (Whittaker, 2021). This study follows actors' definitions of AI, considering how the authors of the documents understand and present AI through the definitions and examples they use. According to Collett and Dillon (2019, p. 7), AI refers to "a heterogeneous network of technologies—including machine learning, natural language processing, expert systems, deep learning, computer vision, robotics—which share in common the automation of functions of the human brain." Similarly, Young et al. (2021) use a definition that refers to AI when a machine or system performs tasks that would ordinarily require human or other biological brainpower to be accomplished. The UNESCO (2020, p. 4) report explains:

> Simply put, artificial intelligence (AI) involves using computers to classify, analyze, and draw predictions from data sets using a set of rules called algorithms. AI algorithms are trained using large datasets so that they can identify patterns, make predictions, recommend actions, and figure out what to do in unfamiliar situations, learning from new data and thus improving over time.

The reports mention numerous examples of AI applications. For example, the UNESCO (2020, p. 4) report says that "we interact with AI on a daily basis in our professional and personal lives, in areas such as job recruitment and being approved for a bank loan, in medical diagnoses, and much more." Through various examples of AI applications reinforcing sexist and racist stereotypes—sentencing algorithms, hiring tools, voice assistants, humanoid robots, chatbots, linguistic biases—a more contested and problematic nature of AI is revealed. These examples highlight that AI is not just a neutral tool but is co-created with society, and as such has major political and social implications in reinforcing existing power relationships, discrimination, and structural inequalities (Benjamin, 2019; Noble, 2018; Ulnicane & Aden, 2023).

To examine how AI impacts diverse social groups, intersectionality with its focus on multiple and overlapping identities provides a useful perspective. While gender has long dominated (and still often dominates) efforts toward equality, intersectionality helps to emphasize and remind us of the importance of addressing multiple inequalities and their reinforcement (Crenshaw, 1991; Fothergill et al., 2019; Verloo, 2006). These multiple identities relate to social categories like gender, race/ethnicity, sexual orientation, and class. Such categories are far from straightforward. For example, the traditional understanding of gender as binary (male/female) is being challenged today (D'Ignazio & Klein, 2020). Examining the convergence of various social identity categories is crucial for understanding processes of oppression, dominance, and power. Regarding the socially constructed categories of identity, Crenshaw (1991, pp. 1296–1297), in her groundbreaking work on intersectionality, highlights that:

> A large and continuing project for subordinated people…is thinking about the way power has clustered around certain categories and is exercised against others. This project attempts to unveil the processes of subordination and the various ways those processes are experienced by people who are subordinated and people who are privileged by them. It is, then, a project that presumes that categories have meaning and consequences. And this project's most pressing problem, in many if not most cases, is not the existence of the categories, but rather the particular values attached to them and the way those values foster and create social hierarchies.

Crenshaw's focus on intersecting identity categories and their relation to power and oppression is highly relevant for AI, where similar patterns of subordination and marginalization have been observed (Allhutter et al., 2020; Benjamin, 2019; Broussard, 2023; Browne et al., 2023; D'Ignazio & Klein, 2020; Little & Winch, 2021; Noble, 2018; Søraa, 2023; Young et al., 2023). Feminist, intersectional, and decolonial approaches to AI present resistance to discriminatory AI, draw attention to the wealth of experience that has been erased from the history of AI (women, trans people, people of color, and the disability community), and call for new forms of solidarity across diverse experiences and identities (Bentley et al., 2023; Ciston, 2019; Png, 2022; Toupin, 2023; S. M. West, 2020).

2.2. Methods and Data

To examine the emerging agenda on intersectionality in AI, this study analyzes four high-profile reports dedicated to intersectionality and AI (Table 1) published between 2019 and 2021. All selected reports explicitly deal with AI rather than digital technology more broadly (e.g., M. West et al., 2019). While some of the selected documents primarily focus on gender, they all explicitly recognize and discuss the importance of an intersectional approach. All four reports, ranging from 33–60 pages, provide an in-depth analysis of the topic. By undertaking a close analysis of a small number of dedicated in-depth reports, this article complements the studies that have interrogated gender and bias issues in AI policy documents at the national and international levels (Guevara-Gómez et al., 2021; Rönnblom et al., 2023; Ulnicane & Aden, 2023).

Three reports are published by well-known AI research centers in the UK and the US: the AI Now Institute, the Leverhulme Centre for the Future of Intelligence, and the Alan Turing Institute. The remaining document (UNESCO, 2020) has been prepared by an international organization. While all four are well-known organizations, they operate quite differently.

The AI Now Institute was founded in 2017 at New York University by internal tech industry critics and was initially funded by tech companies (Sadowski & Phan, 2022). It has regularly prepared influential critical reports on AI, which have been used by government agencies, experts, professional associations, and

Table 1. Overview of the reports analyzed.

Authors	Title	Organization	Date of publication	Number of pages
S. M. West, M. Whittaker, and K. Crawford	*Discriminating Systems: Gender, Race, and Power in AI*	AI Now Institute (US) https://ainowinstitute.org	April 2019	33
C. Collet and D. Dillon	*AI and Gender: Four Proposals for Future Research*	Leverhulme Centre for the Future of Intelligence (UK) http://lcfi.ac.uk	June 2019	43
UNESCO	*Artificial Intelligence and Gender Equality: Key Findings of UNESCO's Global Dialogue*	UNESCO https://www.unesco.org/en	August 2020	49
E. Young, J. Wajcman, and L. Sprejer	*Where Are the Women? Mapping the Gender Job Gap in AI*	The Alan Turing Institute (UK) https://www.turing.ac.uk	2021	60

international organizations (Ulnicane & Aden, 2023). As of 2022, AI Now Institute is an independent organization and currently does not take funding from corporate donors. The report analyzed (S. M. West et al., 2019) was supported by Pivotal Ventures, founded by Melinda Gates. The Leverhulme Centre for the Future of Intelligence at the University of Cambridge is funded by the UK research funding charity the Leverhulme Trust. It is an interdisciplinary AI research center working on a wide variety of research programs on topics such as AI governance, trust, and narratives. The report studied (Collett & Dillon, 2019) was sponsored by the Ada Lovelace Institute—an independent research body—and supported by the consulting company PwC. The Alan Turing Institute is the UK's national institute for data science and AI, launched by leading UK universities and a research council. The report examined (Young et al., 2021) has been prepared by the Women in Data Science and AI project of the Public Policy Program of the institute. The final report (UNESCO, 2020) comes from the United Nations Educational, Scientific and Cultural Organization, based in Paris. It was prepared by the UNESCO Division for Gender Equality and funded by the German Government.

While the organizational and funding contexts from which the four reports originate are diverse, all reports present a critical analysis of gender and diversity in AI. Importantly, all reports are authored by women at various career stages (an exception is the UNESCO report that does not name individual authors), including leading experts in AI like feminist scholar Judy Wajcman, or researchers of social and political aspects of AI, such as Meredith Whittaker and Kate Crawford. From an intersectional perspective, it is important to mention that all four well-known and influential reports originate from a small number of elite institutions in the Global North.

To analyze how these reports define the problems of intersectionality in AI and the ways to address them, this study uses a framing approach (Bacchi, 2000; Rein & Schon, 1996; van Hulst & Yanow, 2016). Framing offers a politically nuanced and power-sensitive way to analyze narratives about the problems and possible ways to act upon them. Its focus is on sense-making, selecting, naming, categorizing, and storytelling in complex situations. This approach has been used to analyze policies relevant to this study such as gender equality (Verloo, 2005) and AI (Ulnicane, 2022). Thus, framing is a highly appropriate approach for this examination of how the selected documents make sense and tell stories about the issues of intersectionality in AI and formulate recommendations for tackling them.

3. Analysis of the Reports on Intersectionality and AI

This section examines the selected documents by focusing on how they frame concerns (Section 3.1), what methods and data they use (Section 3.2), how they approach the various social categories they analyze (Section 3.3), and what recommendations they suggest (3.4). For an overview see Table 2.

3.1. Concerns

The reports raise several interconnected concerns, such as the lack of diversity in the AI sector, reproduction and reinforcement of stereotypes and bias in AI systems, the disproportionate effect of job replacement for women, as well as broader issues of social and economic justice and concentration of power. While previously the diversity problem of the AI industry and issues of bias in AI systems have been examined separately, these reports show that "issues of discrimination in the workforce and in system building are deeply intertwined" (S. M. West et al., 2019, p. 6).

Table 2. Intersectionality in the reports analyzed.

Report	Concerns	Methods and data	Social categories	Examples	Recommendations
S. M. West et al. (2019)	Diversity crisis in the AI sector intertwined with bias and discrimination in AI systems Pushback against diversity	Literature review Data on diversity, pay, discrimination, and harassment	Gender (non-binary, fluid), race, and other identities in the context of existing power structures	Sentencing algorithms Amazon hiring tool Targeted ads Facial recognition	Improving workplace diversity and addressing bias and discrimination in AI systems Worker-led initiatives
Collett and Dillon (2019)	Biased datasets Lack of diversity in the AI workforce	Workshop Literature review	Gender (non-binary) in relation to power dynamics, and functioning intersectionally with race, ethnicity, and sexuality	Humanoid robotics Virtual personal assistants Facial recognition Crime and policing technologies Linguistic biases Health technologies	Four research proposals: bridging gender theory and AI practice; law and policy; biased datasets; diversity in the AI workforce
UNESCO (2020)	AI reinforces gender stereotypes Women at higher risk of being replaced by automation	Dialogue with experts Additional research (including on AI principles)	Gender (non-binary) Intersectionality	Digital voice assistants Amazon hiring tool Targeted ads	Gender equality in AI principles Increasing awareness, education, and skills Industry action Coalition building
Young et al. (2021)	The absence of women in AI and data science leads to gender bias being built into machine learning systems Social and economic justice	Dataset on AI and data science workforce Review of existing datasets Case study of AI platforms Literature review	Gender Binary gender data and lack of intersectional data as a problem	Amazon hiring tool Racist and sexist chatbot Targeted ads Facial recognition Linguistic biases Voice assistants	Reporting standards on gender and intersectional data Government programs Targets and quotas for recruiting and promoting women

Lack of diversity in the AI workforce is seen as the key problem in all four reports. It is called the "diversity crisis" and "diversity disaster" (S. M. West et al., 2019) and illustrated with data on the lack of women, black people, and minorities in the AI sector. Women comprise 15% of AI research staff at Facebook and just 10% at Google, while only 2.5% of the workforce at Google and 4% of the workforce at Facebook and Microsoft is black (S. M. West et al., 2019, p. 5). Furthermore, Facebook reported just 5% of Hispanic workers, and only 3.6% of Google's full-time workers are Latinx (S. M. West et al., 2019, p. 11). Merely 18% of authors at leading AI conferences are women and more than 80% of AI professors are male (S. M. West et al., 2019). Only 7% of

students studying computer science and 17% of those working in technology in the UK are women (Collett & Dillon, 2019, p. 25). Worldwide, 78% of AI and data science professionals are male (Young et al., 2021, p. 2). In the broader field of computer science, women make up 24.4% of the workforce and receive median salaries that are only 66% of the salaries of their male counterparts (S. M. West et al., 2019, p. 11). When discussing the diversity crisis, the reports mention issues of harassment, discrimination, unfair compensation, microaggressions, unwelcoming environments, stereotypes, a culture of inequity, and a lack of promotion for women at tech companies (S. M. West et al., 2019; Young et al., 2021).

Young et al. (2021) find diverging career trajectories, where women are more likely to occupy jobs associated with less status and pay, usually within analytics, data preparation, and exploration, rather than the more prestigious jobs in engineering and machine learning. Moreover, there are fewer women in industries that traditionally entail more technical skills and, thus, are seen as "masculine" like information technology, while more women are in industries that involve fewer technical skills and are perceived as "feminine," such as healthcare. This pattern reflects the historical association of technology with men, while perceiving femininity as incompatible with technology (Wajcman, 2010). Young et al. (2021) find that there are also fewer women in leadership positions in AI, even though they are better qualified, i.e., have higher education levels than men. Still, women self-report fewer skills on LinkedIn than men and are less active on online data science platforms. Furthermore, women working in AI and data science have higher turnover and are more likely to leave the sector than men.

Considering that these issues have been known and investments have been made to address them for decades, the current situation is described as "alarming" (S. M. West et al., 2019). Despite diversity initiatives, the broader field of computer science has experienced a sharp decline of women in its ranks from 37% of computer science majors in the US in 1984 to 18% in 2015 (S. M. West et al., 2019, p. 11). This is reflected in broader concerns discussed in the literature that diversity initiatives have changed little in the AI sector, are still often poorly understood among tech workers, and are sometimes even seen as a threat to scientific excellence (Browne et al., 2024; Stinson & Vlaad, 2024). Furthermore, experts also highlight the pushback to diversity by those who question or even reject that racism, misogyny, and harassment are problems in the AI field (Collett & Dillon, 2019; S. M. West et al., 2019). S. M. West et al. (2019) are critical of so-called "pipeline studies" that focus on the absence of diverse candidates in the hiring pool, which is often used by the industry to justify the lack of diversity and place the problem outside their remit. They refer to the well-known work of historian Mar Hicks who has demonstrated structural discrimination in the computing sector that led to the exclusion and marginalization of women who initially dominated computer operation and programming (Hicks, 2018).

Importantly, the reports emphasize that these are systemic issues reflecting existing power relationships in the AI sector, which are shaped by a feedback loop of discriminatory workplace practices leading to discriminatory tools (S. M. West et al., 2019). The major gender disparity in the AI workforce hinders the development of equitable AI (Collett & Dillon, 2019). "A troubling and persistent absence of women" (Young et al., 2021, p. 2) is seen as having wider consequences when it results in gender bias being built into machine learning systems. It is also seen as fundamentally an ethical issue of social and economic justice. The reports mention numerous cases where AI applications reinforce gender and racial stereotypes and discrimination, including targeted ads shown according to gender and racial stereotypes, or Amazon hiring tool that downgraded women applicants because it was built using historical—predominately

male—employment data (S. M. West et al., 2019). S. M. West et al. (2019, p.5) highlight the systemic nature of the problems:

> The diversity problem is not just about women. It is about gender, race, and most fundamentally about power. It affects how AI companies work, what products get built, who they are designed to serve, and who benefits from their development.

Experts stress that AI tools spreading and reinforcing gender stereotypes would further stigmatize and marginalize women in economic, political, and social life as well as offset an earlier progress made toward gender equality (UNESCO, 2020). They state that the design and implementation of AI perpetuates a vicious cycle, where technology captures and amplifies controlling and restrictive conceptions of gender and race, such as gender binary and racial hierarchies, which are then repetitively reinforced (Collett & Dillon, 2019). This can be seen, for example, in digital voice assistants. Almost all voice assistants, such as Amazon's Alexa or Apple's Siri, "are given female names and voices, and express a 'personality' that is engineered to be uniformly subservient" (UNESCO, 2020, p. 2).

Not only women are largely underrepresented in the AI sector, but they are also at a higher risk of being displaced in the workforce because they are disproportionately represented in sectors that are undergoing automation, like clerical, administrative, bookkeeping, and cashier jobs (UNESCO, 2020).

The reports point out the limitations of emerging AI policies and guidelines. Collett and Dillon (2019, p. 13) emphasize the risk "that economic prosperity and political power will play an underlying role in shaping laws and policies on AI at the expense of more socially equalizing motivations." Analysis of gender equality issues in AI principles reveals that "direct references to gender equality and women's empowerment in existing AI and ethics principles are scarce" (UNESCO, 2020, p. 9). Nevertheless, the report finds implicit alignment with gender equality in discussions of issues such as justice and solidarity.

To sum up, the reports diagnose the diversity situation in the AI field as alarming, troubling, and being in crisis. Problems are seen as systemic, leading to a negative feedback loop and a vicious cycle where discrimination in the workforce leads to the building of discriminating tools. Importantly, experts highlight that previous diversity initiatives have not led to positive changes but rather to a decline in diversity. Despite the acceptance of the diversity rhetoric by tech companies, it is often poorly understood and has experienced some pushback. Reports highlight that AI reinforces gender and racial stereotypes and can offset earlier progress made toward equality.

3.2. Methods and Data

To identify problems and come up with recommendations, the reports draw on a range of research methods and data sources, including a literature review, data on diversity, pay, discrimination, and harassment, workshop and dialogue with experts, and analysis of AI principles. Reports build on feminist and occasionally also intersectional scholarship and review existing literature in this area, including many well-known studies (e.g., Angwin et al., 2016; Benjamin, 2019; Broussard, 2018; Buolamwini & Gebru, 2018; Caliskan et al., 2017; Criado Perez, 2019; D'Ignazio & Klein, 2020; Eubanks, 2019; Hicks, 2018; Noble, 2018; O'Neil, 2016; Zou & Schiebinger, 2018). A considerable part of this literature is authored by women, including black

women. Furthermore, documents refer to each other, with the UNESCO (2020) report citing earlier documents on AI, gender, and race (Collett & Dillon, 2019; S. M. West et al., 2019) and Young et al. (2021) reviewing the reports by UNESCO (2020) and S. M. West et al. (2019). They also review other related reports and guidelines; for example, the UNESCO (2020) document draws on UNESCO's work on digital skills and gender divide (M. West et al., 2019) and AI principles.

Additionally, the reports build on an expert workshop and consultation. One of the documents (Collett & Dillon, 2019) results from the trans-disciplinary and trans-sectoral Gender and AI workshop held at the University of Cambridge. The workshop identified four proposals for future research on bridging gender theory and AI practice, law and policy, biased datasets, and diversity in the AI workforce. It is primarily situated in the UK context but argues that research should be as international as possible (Collett & Dillon, 2019). The workshop had 47 attendees, including well-known experts in the field of AI from industry, government, non-governmental sector, and leading universities like Cambridge, Oxford, and the London School of Economics. Many of them come from the UK but some also from other countries like the US and Ireland.

For the UNESCO report, with a particular focus on the private sector, 17 experts in AI, digital technology, and gender equality from academia, civil society, and the private sector were consulted (UNESCO, 2020). Many of these experts come from the UK and the US. Two out of 17 experts have also attended the workshop that led to the report by Collett and Dillon (2019). The purpose of the dialogue was to identify issues, challenges, and good practices to help overcome the built-in gender biases in AI devices, data sets and algorithms, improve the global representation of women in technical roles and boardrooms in the technology sector, and create robust and gender-inclusive AI principles, guidelines, and codes of ethics within the industry.

Overall, for these two reports (Collett & Dillon, 2019; UNESCO, 2020) a relatively small number of well-known experts have been involved, with many of them coming from only two countries—the UK and the US. From an intersectional perspective, this raises important questions about who has an opportunity to be part of identifying problems and developing recommendations and who is left out.

To study the gender job gap in AI, Young et al. (2021) reviewed existing statistics and datasets in this area. They conclude that existing data is sparse, fragmented, incomplete, and inadequate to analyze the careers of women and men in AI. Shortcomings of available data include binary gender data and the lack of intersectional data on age, race, geography, disability, sexual orientation, and socioeconomic status. The limited data availability hinders opportunities for intersectional analysis (see Section 3.3). In partnership with an executive search and consulting firm specializing in data science, advanced analytics, and AI, they develop a novel data science and AI career dataset. It consists of 19,535 profiles out of which 11.3% are women, belonging mostly to the US, France, Germany, and the UK (Young et al., 2021, 2023). Their sample is very senior with an average of almost 20 years of work experience and with over 55% holding a graduate or postgraduate degree. The authors admit that their sample is not comprehensive, intersectional, and representative of the entire global data and AI population. Thus, although a novel dataset contributes to better information about the little-studied gender issues in AI and data science, it still is limited to senior professionals in a small number of most developed countries.

To summarize, while all four reports use methods and data typical for preparing such documents, like literature review, workshop, dialogue, and a new data set, they have some limitations, especially concerning

intersectional analysis. The sample in the dataset and participants in a workshop and dialogue come predominantly from elite ranks in a tiny number of Global North countries (US, UK, France, and Germany). In a certain way, this selection reproduces existing power imbalances, where those based in the periphery or Global South, and doing less prestigious work in AI and data science, like data labeling or content moderation, remain voiceless and marginalized. The reports admit that their analysis is not comprehensive. The UNESCO (2020, p. 9) report states upfront that "this is not a comprehensive exploration of the complexities of the AI ecosystem in all its manifestations and all its intersections with gender equality. Rather, this is a starting point for conversation and action."

3.3. Social Categories

While one of the reports already in its title explicitly refers to gender and race (S. M. West et al., 2019), the other three documents (Collett & Dillon, 2019; UNESCO, 2020; Young et al., 2021) primarily focus on gender but recognize the importance of intersectional approach. Gender is typically understood as non-binary and like other social categories, it is analyzed in the context of existing power structures. Collett and Dillon (2019), who mainly discuss gender, recognize that gender is inseparable from race, ethnicity, and sexuality. In their report, "gender is understood to have an inextricable relationship with unequal power dynamics, and to function intersectionally with other protected characteristics such as race, ethnicity, and sexuality" (Collett & Dillon, 2019, p. 7). Their focus on equality includes trans, queer, and non-binary equality. Similarly, the UNESCO report, which predominantly explores gender, recognizes the importance of intersectionality:

> Women are a multifaceted and heterogenous group and have different experiences based on realities or characteristics which include: women living in rural and remote areas; indigenous women; racial, ethnic or religious minority women; women living with disabilities; women living with HIV/AIDS; women with diverse sexual orientations and gender identities; younger or older women; migrant, refugee or internally displaced women or women in humanitarian settings. (UNESCO, 2020, pp. 7–8)

Documents discuss a well-known problem (see D'Ignazio & Klein, 2020), namely, the limited availability of intersectional data, such as the lack of data on trans persons and other gender minorities. They find that "the literature almost solely looked at gender, and represented gender as binary. It much less frequently examined race, or other identities, and even more rarely examined the intersection of such identities" (S. M. West et al., 2019, p. 20). S. M. West et al. (2019) emphasize that, instead of the overwhelming and too narrow focus on "women in tech" likely to privilege white women, it is important to acknowledge intersections of gender, race, and other identities as well as existing power structures.

3.4. Recommendations

In response to the concerns discussed, the reports outline several recommendations for tackling the diversity problem in the AI sector as well as improving AI policy and research in this area. Interestingly, one of the reports highlights that its proposals are not intended to be prescriptive but rather should be seen as a provocative mechanism to raise awareness, summarize the current challenges, and prompt practical action (Collett & Dillon, 2019).

Importantly, the UNESCO report emphasizes that to address diversity issues an adequate framing of the overarching AI landscape is needed. It suggests to "shift the narrative of AI as something 'external' or technologically deterministic, to something 'human' that is not happening to us but is created, directed, controlled by human beings and reflective of society" (UNESCO, 2020, p. 33).

The reports emphasize a critical moment and urgency to address the diversity crisis, where "the AI sector needs a profound shift in how it addresses the current diversity crisis. The AI industry needs to acknowledge the gravity of its diversity problem, and admit that existing methods have failed" (S. M. West et al., 2019, p. 3). Recommendations for increasing the diversity of the AI workforce emphasize the need to go beyond just hiring more women and minorities. Collett and Dillon (2019) suggest looking not only at balancing the numbers of the AI workforce but also at creating a culture of diversity in educational institutions and the workplace. When discussing diverse development teams, the UNESCO report argues for a broad approach emphasizing that "this is not a matter of numbers, but also a matter of culture and power, with women actually having the ability to exert influence" (UNESCO, 2020, p. 23). Additionally, it calls for a robust approach to raise awareness and literacy, technical and ethical education, skills development, and capacity building.

Similarly, S. M. West et al. (2019) underline the systemic nature of problems of diversity and bias in AI, which cannot just be addressed by technical approaches but rather require an integration of social and technical approaches. Recommendations to address diversity problems include ending pay and opportunity inequality, developing and implementing reporting standards on gender and other characteristics in AI and data science companies, publishing transparency reports on harassment and discrimination, and broadening recruitment beyond elite universities. Young et al. (2021) call for intersectional gender mainstreaming in human resources policy so that women and men are given equal access to well-paid jobs and careers as well as for actionable targets, incentives, and quotas for recruiting and promoting women. S. M. West et al. (2019) argue that a broader approach is needed that considers existing racism and misogyny and focuses on changing workplace cultures. They highlight the role of worker-led initiatives to address the issues of harassment, discrimination, diversity, and equal pay.

To tackle the development of biased and discriminatory AI tools, experts call for going beyond technical debiasing to include a wider social analysis (S. M. West et al., 2019). Reports stress the importance of diverse perspectives and participation of a wide range of stakeholders from academia, civil society, government, and the private sectors. Such multi-stakeholder forums would include users as well as diverse disciplines from computing to social sciences and humanities. The UNESCO report suggests "establish[ing] a multi-disciplinary and inter-generational coalition that builds partnerships across sectors and groups for a holistic society/AI ecosystem approach. Create avenues for dialogue and learning by creating a common understanding and language and through collaborations and collective impact models" (UNESCO, 2020, p. 34). However, Young et al. (2021) emphasize that the AI industry should avoid "participation washing," when the mere fact that somebody has participated in a project (rather than having had a chance to meaningfully shape it) is supposed to lend it legitimacy.

The UNESCO (2020) report mentions that AI has the potential to contribute to gender equality. To illustrate this point, the report provides an example that AI can help employers use gender-sensitive language to write inclusive job postings to increase the diversity of their workforce. In the reports studied, this is a rare example of the positive impact of AI on equality.

Importantly, the reports also suggest exploring where AI should not be deployed (Collett & Dillon, 2019), if it results in inequality and discrimination. Experts recommend undertaking risk assessments of whether certain systems should be designed at all (S. M. West et al., 2019). Tools that claim to detect sexuality from headshots, predict criminality based on facial features, or assess worker competence via micro-expressions are seen as particularly problematic and in need of urgent reconsideration (S. M. West et al., 2019). To avoid harm, the UNESCO report suggests accepting "that some things may not be able to be fixed and therefore should not be done at all, or should ultimately be abandoned (e.g., the example of Amazon's hiring algorithm which remained biased after multiple attempts to fix it)" (UNESCO, 2020, p. 17).

To address diversity issues in AI, reports emphasize the need for effective and gender-inclusive policies (Young et al., 2021). The UNESCO (2020) report calls for governments to commit to policies, regulations, and mechanisms that proactively promote gender equality in and through AI. Experts suggest that governments together with national and international organizations must initiate research and advocacy programmes to tackle gender gaps. To ensure justice and fairness, they should also take proactive steps to include women and marginalized groups in the design and development of AI (Young et al., 2021). Moreover, the reports focus on guidelines and principles that can support diversity. The UNESCO report suggests strengthening and operationalizing gender equality in AI principles, while Collett and Dillon (2019) recommend context and gender-specific guidelines for data collection and handling, in particular in three contexts that significantly impact gender equality: crime and policing, health, and the financial sector.

Experts call for research in this area to be collaborative, intersectional, pluralistic, interdisciplinary, and trans-sectoral (Collett & Dillon, 2019). It needs to provide a deeper analysis of the power and structural challenges AI systems pose to communities examining the relationship between technology development and the lived experiences of individuals of different racialized, gendered, and classed identities (S. M. West et al., 2019). Collett and Dillon (2019) suggest utilizing gender theory, including trans, non-binary, queer, feminist, as well as postcolonial theory. Considering the crucial role that law and policy play in shaping AI, they recommend analyzing their impact on gender, as well as race, ethnicity, sexuality, disability, etc., so that "the intersectional nature of the research would enable it to consider different standpoints, working for widespread social justice and redistribution of power" (Collett & Dillon, 2019, p. 18). The UNESCO (2020) report highlights the importance of learning from past experiences: successes, gaps, and failures of diversity initiatives.

To summarize, the reports emphasize the urgency and critical moment to acknowledge the gravity of the diversity problem in AI. They call for a holistic and broad approach that goes beyond just increasing the numbers of women in AI and focuses on culture, power, and opportunities to exert influence. Rather than just focusing on technical fixes of bias, a socio-technical approach that involves perspectives from multiple disciplines and sectors is suggested (see Ulnicane & Aden, 2023). Importantly, the reports argue that some AI systems, that result in discrimination and inequality, should not be built at all.

4. Conclusions

While the political agenda on AI tends to be dominated by questions of potential economic and social benefits (Schiff, 2023; Ulnicane, 2022), the four high-profile reports on intersectionality in AI analyzed in this study play an important role in bringing the problematic impacts of AI on gender, race, socio-economic background, and other intersecting identities to broader attention.

This study demonstrates that there is a lot of convergence among the four reports in the way they frame problems, provide evidence through data and examples, and outline recommendations for action (Table 2). Experts demonstrate how the diversity crisis in the AI workforce is closely intertwined with biased AI tools. They emphasize the systemic nature of problems when discrimination in the workforce leads to the development of discriminatory systems. This diagnosis of the problem resonates with feminist scholarship of technology that focuses on the mutual shaping of gender and technology and highlights how processes of technical change can influence gender power relations (Wajcman, 2010). A major concern discussed in the reports is the ineffectiveness of earlier diversity initiatives in tech. The reports highlight the urgency and critical moment to reconsider previous approaches and address problems in a holistic way that focuses not only on increasing the number of women and minorities but also on considering culture, power, and opportunities to exert influence.

Remarkably, almost all examples of the impact of AI on gender, race, and other identities in these reports as well as in other AI documents (Ulnicane & Aden, 2023) highlight problematic consequences, while it is difficult to find examples that would confirm optimistic views that AI could help to eliminate human bias. In the four reports analyzed, the only positive example suggests that AI could help to use gender-sensitive language to write inclusive job postings to help increase the diversity of their workforce (UNESCO, 2020).

These reports, written by some of the leading scholars on feminism and the social implications of AI, are important in highlighting the AI diversity problems. However, from an intersectional perspective, these influential reports originating from a small number of elite institutions in the Global North are somewhat limited because their data and methods mainly focus on senior professionals in a few most developed countries.

Nevertheless, the problems and recommendations outlined in these reports are important initial steps in shaping the agenda and calling for action on intersectionality in AI. These reports, and most of the literature they cite, come from a community of women dedicated to the issues of intersectionality and AI. However, in the patriarchal system, these issues can be at risk of being pigeonholed as issues relevant mostly to women, black people, and minorities rather than part of the core AI agenda (D'Ignazio & Klein, 2020; Ulnicane & Aden, 2023). Thus, one of the questions for future research is to explore how intersectional and feminist approaches can be seen not just as an add-on but rather as a challenge and alternative to the existing system.

This study contributes to the growing intersectionality research across disciplines and sectors, including AI and data science (Bentley et al., 2023). While further work examining and tackling intersectionality issues in AI is urgently needed, it has to be recognized that intersectionality is not a straightforward concept. Rather:

> [It is] a multifaceted area of theory and praxis that is often contradictory. Intersectionality can be an expression of one's identity, which can be singular, multiple, and/or intersectional, rooted and stable, or changing constantly. It can constitute belonging and/or unpack marginalization and disadvantage. It can unite people in their endeavours and/or detonate struggles against systems of oppression, discrimination or persecution. It can be an abstract ideological project and/or rich and detailed experiences of existence. Crucially, it is a significant forum for investigating and transforming relationships between people, places and institutions, towards human rights, reduced inequality and

> social justice. Likewise, it can be and do none of those things—serving merely as a buzzword. (Bentley et al., 2023, p. 13)

Acknowledgments

Helpful comments and suggestions from two anonymous reviewers are gratefully acknowledged.

Funding

The research reported in this article has received funding from the EU's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3).

Conflict of Interests

The author declares no conflict of interests.

References

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data*, *3*, Article 5. https://doi.org/10.3389/fdata.2020.00005

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *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

Bacchi, C. (2000). Policy as discourse: What does it mean? Where does it get us? *Discourse: Studies in the Cultural Politics of Education*, *21*(1), 45–57.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.

Bentley, C., Muyoya, C., Vannini, S., Oman, S., & Jimenez, A. (2023). Intersectional approaches to data: The importance of an articulation mindset for intersectional data science. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231203667

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Browne, J., Cave, S., Drage, E., & McInerney, K. (Eds.). (2023). *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines*. Oxford University Press.

Browne, J., Drage, E., & McInerney, K. (2024). Tech workers' perspectives on ethical issues in AI development: Foregrounding feminist approaches. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517231221780

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Ciston, S. (2019). Intersectional AI is essential: Polyvocal, multimodal, experimental methods to save artificial intelligence. *Journal of Science and Technology of the Arts*, *11*(2), 3–8. https://doi.org/10.7559/citarj.v11i2.665

Collett, C., & Dillon, D. (2019). *AI and gender: Four proposals for future research*. The Leverhulme Centre for the Future of Intelligence. https://doi.org/10.17863/CAM.41459

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299.

Criado Perez, C. (2019). *Invisible women: Data bias in a world designed for men*. Random House.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

Fothergill, B. T., Knight, W., Stahl, B. C., & Ulnicane, I. (2019). Intersectional observations of the Human Brain Project's approach to sex and gender. *Journal of Information, Communication and Ethics in Society*, *17*(2), 128–144. https://doi.org/10.1108/JICES-11-2018-0091

Guevara-Gómez, A., de Zárate-Alcarazo, L. O., & Criado, J. I. (2021). Feminist perspectives to artificial intelligence: Comparing the policy frames of the European Union and Spain. *Information Polity*, *26*(2), 173–192.

Hicks, M. (2018). *Programmed inequality: How Britain discharged women technologists and lost its edge in computing*. MIT Press.

Little, B., & Winch, A. (2021). *The new patriarchs of digital capitalism: Celebrity tech founders and networks of power*. Routledge.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

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

Png, M.-T. (2022). At the tensions of South and North: Critical roles of Global South stakeholders in AI governance. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1434–1445). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533200

Radu, R. (2021). Steering the governance of artificial intelligence: National strategies in perspective. *Policy and Society*, *40*(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728

Rein, M., & Schon, D. (1996). Frame-critical policy analysis and frame-reflective policy practice. *Knowledge and Policy*, *9*, 85–104. https://doi.org/10.1007/BF02832235

Rönnblom, M., Carlsson, V., & Öjehag-Pettersson, A. (2023). Gender equality in Swedish AI policies. What's the problem represented to be? *Review of Policy Research*, *40*(5), 688–704. https://doi.org/10.1111/ropr.12547

Sadowski, N., & Phan, T. (2022). "Open secrets": An interview with Meredith Whittaker. In T. Phan, J. Goldenfein, D. Kuch, & M. Mann (Eds.), *Economies of virtue: The circulation of "ethics" in AI* (pp. 140–152). Institute of Network Cultures.

Schiff, D. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy. *Review of Policy Research*, *40*(5), 729–756. https://doi.org/10.1111/ropr.12535

Schopmans, H., & Cupac, J. (2021). Engines of patriarchy: Ethical artificial intelligence in times of illiberal backlash politics. *Ethics & International Affairs*, *35*(3), 329–342.

Søraa, R. (2023). *AI for diversity*. CRC Press.

Stinson, C., & Vlaad, S. (2024). A feeling for the algorithm: Diversity, expertise, and artificial intelligence. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517231224247

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Toupin, S. (2023). Shaping feminist artificial intelligence. *New Media & Society*, *26*(1), 580–595. https://doi.org/10.1177/14614448221150776

Ulnicane, I. (2022). Emerging technology for economic competitiveness or societal challenges? Framing purpose in artificial intelligence policy. *Global Public Policy and Governance*, *2*, 326–345. https://doi.org/10.1007/s43508-022-00049-8

Ulnicane, I., & Aden, A. (2023). Power and politics in framing bias in artificial intelligence policy. *Review of Policy Research*, *40*(5), 665–687. https://doi.org/10.1111/ropr.12567

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, *40*(2), 158–177. https://doi.org/10.1080/14494035.2020.1855800

UNESCO. (2020). *Artificial intelligence and gender equality. Key findings of UNESCO's global dialogue*. https://unesdoc.unesco.org/ark:/48223/pf0000374174

van Hulst, M., & Yanow, D. (2016). From policy "frames" to "framing" theorizing a more dynamic, political approach. *The American Review of Public Administration*, *46*(1), 92–112.

Verloo, M. (2005). Mainstreaming gender equality in Europe. A critical frame analysis approach. *The Greek Review of Social Research*, *117*, 11–34. https://doi.org/10.12681/grsr.9555

Verloo, M. (2006). Multiple inequalities, intersectionality and the European Union. *European Journal of Women's Studies*, *13*(3), 211–228.

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics*, *34*(1), 143–152.

West, M., Kraut, R., & Ei Chew, H. (2019). *I'd blush if I could: Closing gender divides in digital skills through education*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

West, S. M. (2020). Redistribution and rekognition: A feminist critique of algorithmic fairness. *Catalyst: Feminism, Theory, Technoscience*, *6*(2), 1–24.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race, and power in AI*. AI Now Institute. https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2

Whittaker, M. (2021). The steep cost of capture. *Interactions*, *28*(6), 50–55. https://doi.org/10.1145/3488666

Young, E., Wajcman, J., & Sprejer, L. (2021). *Where are the women? Mapping the gender job gap in AI*. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/report-where-are-women-mapping-gender-job-gap-ai

Young, E., Wajcman, J., & Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, *38*(3), 391–414.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature*, *559*, 324–326 https://doi.org/10.1038/d41586-018-05707-8

About the Author

Inga Ulnicane is a research fellow at the University of Birmingham (UK) and an honorary senior research fellow at De Montfort University (UK). She works at the intersection of political science and social studies of science and technology. She has published extensively on topics such as politics and policy of AI, governance of emerging technologies, responsible research and innovation, and grand challenges. In addition to her academic research, she has prepared commissioned reports for the European Parliament and the European Commission.

Social Inclusion
2024 • Volume 12 • Commentary 8337
https://doi.org/10.17645/si.8337

cogitatio

COMMENTARY

Open Access Journal

How to Include Artificial Bodies as Citizens

Pramod K. Nayar

Department of English, University of Hyderabad, India

Correspondence: Pramod K. Nayar (pramodknayar@uohyd.ac.in)

Submitted: 4 March 2024 **Accepted:** 23 May 2024 **Published:** 11 July 2024

Issue: This commentary is part of the issue "Artificial Intelligence and Ethnic, Religious, and Gender-Based Discrimination" edited by Derya Ozkul (University of Warwick), fully open access at https://doi.org/10.17645/si.i236

Abstract

This essay ponders on the thorny issue of including artificial beings under the category of "citizen." The increasing humanization of the artificial being, it suggests, prevents us from seeing and treating the machine as a being. But if the humanoid robot performs all the functions of a human being, and acquires cultural traits such as emotional intelligence, rational thinking, or altruism, then on what grounds do we deny it the same status as a human person? Conversely, as more and more humans are cyborged, through transplants, implants, and prostheses, resulting in an erasure of their "core" humanity, then what is the difference between such a cyborged human with human rights and an artificial being?

Keywords

artificial beings; citizenship; humanoid robots; moral standing; personhood

Humanoid robots are now part of healthcare, geriatric care, and childcare. They have been integrated into families in countries like Japan (Robertson, 2018). When the humanoid robot, Sophia, was granted citizenship in 2017 by the Kingdom of Saudi Arabia, it invited the question: Can we include sentient, humanoid robots under the category of "citizens"?

Before we attempt an answer to this question we need to think in terms of how social norms, policies, and institutions (by which I mean state, corporate, technological-industrial, and research institutions) determine the context of the creation, assimilation, and regulation of some technologies rather than others. Ruha Benjamin has termed the "social biases [that] get coded, not only in laws and policies, but in many different objects and tools that we use in everyday life," "discriminatory design" (Benjamin, 2019, p. 5). That is to say, biases and preferences shape the technology that measures, validates, facilitates, or hinders the value of human—or, for that matter, humanoid-android—lives. From facial recognition technology to early crime-prevention technologies and datafication, biases and norms determine any kind of technological

innovation. Safiya Umoja Noble has argued that "digital decisions reinforce oppressive social relationships and enact new modes of racial profiling" and the "people who make these decisions hold all types of values, many of which openly promote racism, sexism, and false notions of meritocracy" (Noble, 2018, pp. 1–2). Whether the human applicant for a job, a credit card, or a loan is deserving of just a machinic evaluation or a human one is built into the algorithm that determines whether the application/applicant meets the qualifying criteria. Such an "algorithmic accountability" embodies the biases of the coders (Broussard, 2018, pp. 43–44).

Robot or bot designs embody a "discriminatory design," because the default option for such bots is white (Poster, 2019). In the words of Neda Atanasoski and Kalindi Vora, "historical forms of domination and power…get built into seemingly non-human objects and the infrastructures that link them, thus sanitizing digital media [and a variety of other] technologies as human-free" (Atanasoski & Vora, 2015, p. 5). The human appearance of the humanoid being, then, is shaped by the desire or fantasy to craft a machine in one's own (human) image. Further, the behaviour of the humanoid robot has to approximate to that of the human.

Yet, it is this approximation of the humanoid to the human that restrains us, I suggest, from opening our arms, metaphorically and literally, to the autonomous machine being employed for care, dangerous tasks or entertainment. That is, the increasing humanization of the robot is itself discriminatory because of a social imaginary that prevents us from treating the machine as a *being*. Its humanity, we remind ourselves, is programmed: it is a non-human object that merely mimics the human. And yet, the non-human is created with the task of providing or performing the hitherto human tasks of domestic work, care, exploration (robots sent in ahead of humans to check the terrain), companionship. In other words, the very design of a humanoid robot is discriminatory because it alerts us to the co-existence in the same "person" (of the robot) of what we traditionally take to be incompatible features: the machine and the human. The human's profile of the robot or artificial being is speciesist: the robot is a different species, although in the posthuman age we know that millions of humans have incorporated, from pacemakers to more advanced technological devices, machines into their organic body. Discriminatory design is the merger of the biotic and the abiotic in the personhood of the artificial being.

Considering the artificial being or robot as a person means embarking on a process of radical social inclusivity—the social now involving, literally, the humanoid robot and, historically, the animal—depends on answers to a series of conflicted and often confounding but interrelated questions. These questions are centered around the moral standing and personhood of humanoid robots for, in the words of Rosi Braidotti, "only ethical and legal issues remain to be solved to grant responsibility to autonomous machines' decision making, while the cognitive capacities are already in place" (Braidotti, 2013, p. 44).

If the robot can undertake tasks such as care, exhibit rational thinking, emotional intelligence, even biases, and generally function as humans do, then on what grounds can we deny them inclusion in the category "citizen"? That is, if their actions are analogous to those of humans, and if such actions by humans would automatically result in an obligation to those who perform those (as in a care relation), then is it not possible to think of a robot performing care relations producing an obligation towards "it"? (LaBossiere, 2017; for "care robot" definitions see Vallor, 2011; van Wynsberghe, 2013). Would it not be, to phrase it differently, discriminatory to say, "despite its appearance, behaviour, skills and functions, we are not obligated to the carer robot"?

If the argument is that robot emotional intelligence is programmed and not "natural" to "it," then this argument is inherently flawed because psychopaths, people with brain injuries and other conditions do not demonstrate the same emotional responses as normative humans do. Conversely, we *acquire* emotional intelligence through cultural training where, for instance, we learn to present specific kinds of appropriate emotional responses to the events and people around us. So, if cultural training induces emotional intelligence in humans, why is the algorithmic and generative emotional intelligence of the artificial being unacceptable?

If the humanoid robot acquires its skills and cultivates its potentialities within the human social order, would "it" not be on par with the human? With AI, they learn to traverse human dynamics, as science fiction and dystopian novels such as Ian McEwan's *Machines Like Me* (2019) and Kazuo Ishiguro's *Klara and the Sun* (2021) portray. Humans reach and fulfil their potential within a socius and set of relationships. Our identities emerge from these sets of relations, which include care, nurture, support. Robots introduced as carers, companions, and even as servitors integrate into households, families, and the general social order. In such a context, would the artificial body not be entitled to the same status as a human? If, human rights laws apply to all human persons and prohibit slavery, then is not the servitude of beings created as humanoids and undertaking servitor work for humans covered by the ambit of such laws (Nayar, 2023)?

If the human's social dynamics already involve the non-human, such as pets, seeing-eye dogs, comfort creatures, and others, the animal is seen as a "member" of the human family although, at times, the creature behaves like a *creature* (Fudge, 2002). Just as a human's location within social categories—class, race, ethnicity, gender, work/profession—determines how we evaluate her/his worth, it is the robot's insertion into the social order (family robot, carer robot, military robot) that determines its worth for humans and whether we accept or discriminate against them. Further, if humans are accepted based on their cultural markers—religion, ethnicity, language—then this applies to robots as well.

As we expand to include service robots, care robots, and social robots (companion robots) in the socius, would they not be, like pets or seeing-eye dogs, a part of the dynamic? If the humanoid robots are designed to "provid[e] a sense of companionship and intimacy similar to the intimacy provided by companion animals," then how are they categorically different from the animal? (DeFalco, 2023, pp. 31–32). Is "real" care the province solely of the creaturely and not of the artificial, DeFalco queries?

If human dynamics, especially of care and affect, are premised upon the "more immediate connection at the heart of cultural imaginaries of affection," as Puig de la Bellacasa puts it (2017, p. 103), then is the creaturely touch and connection so radically distinct from the equally *touchable* connection offered by care robots? Further, if *human* sociality, especially in the elderly, the chronically ill, the socially inept, hinges on the carer robot's companionship—the human–non-human dynamics—then on what grounds do we discriminate against the robot by excluding it from the status of "persons"?

If the human's evolution has been, as commentators (Braidotti, 2013) suggest, a *co*-evolution with both technology and the non-human, then should kinship not be seen as more-than and other-than-human as well? The domesticated animal was a part of human evolutionary and civilizational history, and while it was not "family" it was a sort of kin. Conversely, some classes of humans—slaves, for example—were never seen as either kin or family, while horses, dogs, and others were treated like members of the family. So now when we have the non-human-but-humanoid robot as a part of the clan, socius, and household, why would "it" not be kin? Why should we assume that kinship is always *only* human?

If human rights are directed at the human being entitled to fulfill her potential and aspirations, and we now have humanoid robots whose potential is designed and directed at serving the human, then would it be ethical to curb "its" potential? That is, if we have created robots programmed to serve humanity, does it not follow that we should ensure conditions in which "it" fulfills the potential it was made for/with (Petersen, 2007)?

If the evolution of the human is increasingly through techno-pharmacological innovation and intervention, then most humans are already cyborgs and posthuman. With medical and cognitive bioenhancement we see extended health spans and more immunological traits. We also see signs of moral bioenhancement that can not only erase "deviant" tendencies but actually amplify the more socially acceptable and valued qualities like altruism (Buchanan, 2011). In these contexts, as we cyborgise or posthumanize large sections of humanity, where do we draw the line of the humanity of the cyborg itself?

If the human aim is to become, collectively, less vulnerable to danger (including disease, injury, and mortality, especially in the transhumanism propounded by Nick Bostrom and others), then this same cyborgization will produce the enhanced-and-less-vulnerable as compared to the non-enhanced human persons. This means new classes of vulnerable humans will emerge. In such a context, is the vulnerability of the humanoid robot to danger any different from the vulnerability of the non-enhanced human, for we recognize that vulnerability is not exclusive to the human (Coeckelbergh, 2011; Leido & Rueda, 2021).

If the human is marked by empathy, then we need to distinguish between an embodied affective empathy and cognitive empathic ability. The question, as Stefan Herbrechter enunciates it, is "whether a robot (or software, or smart environments, AI, etc.) *understand* [sic] empathy, whether they can *know* what humans *feel* (which would of course make them virtually human" (Herbrechter, 2022, p. 186, emphasis in original). If humanoid robots possess the latter—the ability to understand and eventually mimic human empathy even if empathy is not an embodied state in "it," then how/why would we see them as less than human?

If the dignity of the human person—enshrined in the Universal Declaration of Human Rights—inheres in her ability to exercise her rationality and the autonomy to make choices (Mendz & Cook, 2021), then is it possible to deny this dignity to humanoid robots, who approximate to the human in several ways and differ from it in several, who are able to do the same?

References

Atanasoski, N., & Vora, K. (2015). Surrogate humanity: Posthuman networks and the (racialized) obsolescence of labor. *Catalyst: Feminism, Theory, Technoscience*, *1*(1) https://catalystjournal.org/index.php/catalyst/article/view/ata_vora

Benjamin, R. (Ed.). (2019). *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life*. Duke University Press.

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

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Buchanan, A. (2011). *Beyond humanity? The ethics of biomedical enhancement*. Oxford University Press.

Coeckelbergh, M. (2011). Vulnerable cyborgs: Learning to live with our dragons. *Journal of Evolution & Technology*, *22*(1), 1–9.

DeFalco, A. (2023). *Curious kin in fictions of posthuman care*. Oxford University Press.

Fudge, E. (2002). *Animal*. Reaktion.

Herbrechter, S. (2022). *Before humanity: Posthumanism and ancestrality*. Brill.

Ishiguro, K. (2021). *Klara and the Sun*. Faber and Faber.

LaBossiere, M. (2017). Testing the moral status of artificial beings; or "I'm going to ask you some questions…" In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 293–306). Oxford University Press.

Leido, B., & Rueda, J. (2021). In defense of posthuman vulnerability. *Scientia et Fides*, *9*(1), 215–239.

McEwan, I. (2019). *Machines like me and people like you*. Jonathan Cape.

Mendz, G. L., & Cook, M. (2021). Posthumanism: creation of 'new men' through technological innovation. *The New Bioethics*, *27*(3), 197–218.

Nayar, P. K. (2023). Artificial beings, servitude and rights: Kazuo Ishiguro's *Klara and the Sun*. In M. Ashraf Raja & N. T. C. Lu (Eds.), *The Routledge companion to literature and social justice* (pp. 515–527). Routledge.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Petersen, S. (2007). The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence*, *17*(1), 43–54.

Poster, W. (2019). Racialized surveillance in the digital service economy. In R. Benjamin (Ed.), *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life* (pp. 133–169). Duke University Press.

Puig de la Bellacasa, M. (2017). *Matters of care: Speculative ethics in more than human worlds*. University of Minnesota Press.

Robertson, J. (2018). *Robo sapiens japanicus*: *Robots, gender, family, and the Japanese nation*. University of California Press.

Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the 21st century. *Philosophy and Technology*, *24*, 251–268.

van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics*, *9*(2), 407–433.

About the Author

Pramod K. Nayar holds the UNESCO Chair in Vulnerability Studies at the Department of English, the University of Hyderabad. Among his newest books are *Vulnerable Earth: The Literature of Climate Crisis* (Cambridge 2024), *Nuclear Cultures: Irradiated Subjects, Aesthetics and Planetary Precarity* (Routledge 2023), *Alzheimer's Disease Memoirs: Poetics of the Forgetting Self* (Springer, 2022), *The Human Rights Graphic Novel*: *Drawing it Just Right* (Routledge 2021), among others, besides essays on human rights, vulnerability, and literature in numerous journals and anthologies.

SOCIAL INCLUSION
ISSN: 2183-2803

Social Inclusion is a peer-reviewed open access journal which provides academics and policymakers with a forum to discuss and promote a more socially inclusive society.

The journal encourages researchers to publish their results on topics concerning social and cultural cohesiveness, marginalized social groups, social stratification, minority-majority interaction, cultural diversity, national identity, and core-periphery relations, while making significant contributions to the understanding and enhancement of social inclusion worldwide.

cogitatio

www.cogitatiopress.com/socialinclusion