Science and Engineering Ethics (2024) 30:46
https://doi.org/10.1007/s11948-024-00507-y

ORIGINAL RESEARCH/SCHOLARSHIP

Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence

Salla Westerstrand[1]

Received: 27 May 2023 / Accepted: 5 August 2024 / Published online: 9 October 2024

Abstract

The popularisation of Artificial Intelligence (AI) technologies has sparked discussion about their ethical implications. This development has forced governmental organisations, NGOs, and private companies to react and draft ethics guidelines for future development of ethical AI systems. Whereas many ethics guidelines address values familiar to ethicists, they seem to lack in ethical justifications. Furthermore, most tend to neglect the impact of AI on democracy, governance, and public deliberation. Existing research suggest, however, that AI can threaten key elements of western democracies that are ethically relevant. In this paper, Rawls's theory of justice is applied to draft a set of guidelines for organisations and policy-makers to guide AI development towards a more ethical direction. The goal is to contribute to the broadening of the discussion on AI ethics by exploring the possibility of constructing AI ethics guidelines that are philosophically justified and take a broader perspective of societal justice. The paper discusses how Rawls's theory of justice as fairness and its key concepts relate to the ongoing developments in AI ethics and gives a proposition of how principles that offer a foundation for operationalising AI ethics in practice could look like if aligned with Rawls's theory of justice as fairness.

Keywords Artificial intelligence · AI ethics · Rawls · Justice as fairness · Democracy

Introduction

The recent popularisation of Artificial Intelligence (AI) technologies has prompted discussion about their ethical implications. Many have speculated over the potential of AI systems to threaten societal structures and quality of life (Bostrom, 2016, 2017; Coeckelbergh, 2022a, 2024; Russell, 2019). This development has pushed organisations to steer the development and application of AI systems towards

✉ Salla Westerstrand
salla.k.westerstrand@utu.fi

[1] Turku School of Economics, Information Systems Sciences, University of Turku, Turku, Finland

a more ethical direction through principles and guidelines (see Jobin et al., 2019; Hagendorff, 2020; Ayling & Chapman, 2022). This trend also known as principalism (Clouser & Gert, 1990) has raised questions about the applicability of these principles in practice (Hickok, 2021; Mittelstadt, 2019), which has led to an increasing volume of research about operationalisation of ethics principles (Bleher & Braun, 2023; Morley et al., 2021, 2023; Stix, 2021). Whereas research around operationalising is essential in order for the principles to effectively guide AI development and deployment, I argue that the work around actionable ethics principles needs revisiting.

Many of the principles in existing guidelines are overlapping (Ashok et al., 2022; Hagendorff, 2020; Hunkenschroer & Luetge, 2022; Jobin et al., 2019), but they also differ in ways that bring forth challenges. They seem to differ in how principles are interpreted (Jobin et al., 2019) and seem to rarely guide the reader through the reasoning leading to the principles (Franzke, 2022; Jobin et al., 2019). Whereas the reviews such as those of Jobin et al., (2019) and Hagendorff (2020) map principles and values found in the guidelines, the very use of *ethics* behind the guidelines remains obscured (Franzke, 2022). In common language, one might talk about ethical questions while referring to predefined rights and wrongs in the given normative context. As Stahl (2022) points out, such issues could be better described as *social concerns* rather than issues of ethics. An ethicist, in contrast, could ask: Why are the suggested principles the most ethical ones? If the guidelines do not offer justifications, are the guidelines truly *ethics* guidelines, or just values or opinions, or a result of a political processes? As Bleher and Braun (2023) put it, if we do not ground our principles into rigorous ethical reasoning, they and the tools for their operationalisation risk being"either inappropriate, meaningless, or merely an end in themselves" (p. 10).

Moreover, Hagendorff (2020) observed that most AI ethics guidelines he reviewed tend to neglect impacts of AI on democracy, governance, and political deliberation. Meanwhile, several studies imply that the current developments in AI might threaten democracy at large (Coeckelbergh, 2022a, 2024) and harm democratic governance by distorting political opinion formation and elections (Alnemr, 2020; Chesney & Citron, 2019; Feezell et al., 2021; König & Wenzelburger, 2020; Manheim & Kaplan, 2019; Nemitz, 2018; Paterson & Hanley, 2020), eroding trust towards democratic institutions (Chesney & Citron, 2019; Manheim & Kaplan, 2019; Paterson & Hanley, 2020), and violating fundamental democratic values, such as equality and justice (Hacker, 2018; Janssen et al., 2022; König & Wenzelburger, 2020; Tolan, 2019). As AI systems with ever broader collective impacts on societies get popularised, such as generative AI systems, a need for shift in perspective from human-in-the-loop towards society-in-the loop suggested by Rahwan (2018) gains ever more relevance. Ethics as an approach seems to have potential for increasing our understanding of the relationship between AI and democracy (Westerstrand, 2023), and Stahl also noted a connection between the ongoing AI ethics discourse with regulation, such as human rights (Stahl, 2022). This implies that to improve the state of AI systems and their alignment with collective values, it might be wise to inspect the larger ecosystem and the ethical foundations behind its guiding principles.

In this paper, I contribute to the discussion through the perspective of John Rawls's theory of justice as fairness. Rawls intended his contractarian theory to build "the most appropriate moral basis for a democratic society" (Rawls, 1971, p. viii), which makes it a theoretically interesting foundation for discussing the ethicality of AI development that comes with an increasing collective ethical and societal implications. By taking a relatively unpopular direction and contributing to the principalist paradigm in contrast to several other applications of Rawls's theory concentrating on algorithmic applications (e.g., Leben, 2017, 2018; Heidari et al., 2019; see also Keeling, 2018), this paper aims to strengthen the ethical rigour of the principalist discourse that keeps informing the technical development of algorithms, the spheres of policy and regulation, as well as strategic decision-making regarding which technologies we should develop and deploy in the first place, and in which contexts.

The paper begins with an overview of Rawls's theory of justice in the context of AI, which serves as a theoretical starting point for revisiting AI ethics guidelines. Then, a suggestion is drafted for a set of ethics guidelines that is based on Rawls's principles of justice as fairness. Finally, conclusions are drawn, the academic and practical potential of the work are discussed, and possibilities for future research and application are identified.

AI Ethics from a Rawlsian Perspective

John Rawls established his theory of justice as fairness mostly in his books *A Theory of Justice* (first published in 1971, revised in 1999), and *Political Liberalism* (1993). He intended his contractarian theory to form the most adequate moral basis for a democratic society (Rawls, 1999, p. 6). He introduced a set of principles he argues would have been agreed upon in the fairest possible setting by neutral, mutually disinterested individuals. According to Rawls, the principles read as follows:

> a) Each person has an equal right to a fully adequate scheme of equal basic liberties which is compatible with a similar scheme of liberties for all. b) Social and economic inequalities are to satisfy two conditions. First, they must be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they must be to the greatest benefit of the least advantaged members of society. (Rawls, 2005, p. 291).

These principles would form the moral basis for the institutions belonging to the basic structure of society, deciding upon basic liberties and the distribution of opportunities and inequalities. Which institutions belong to the basic structure of society is not, however, unambiguous. In academic discourse, some argue that the basic structure only includes legally coercive institutions and not private corporations (*coercive account*), whereas others suggest the defining factor to be the profoundness of the impact of an institution to people's lives (*profound*

effects account) (Berkey, 2021; see, e.g., the profound effects account by Blanc & Al-Amoudi, 2013, and the coercive account of Singer, 2015). Rawls himself only provides partially contradictory statements (e.g., Rawls, 1971, 1999, p. 6 vs. Rawls, 2005, p. 265–266), which opens the discussion for differing interpretations.

In the context of AI, the discussion is further complicated by the variety of organisations in the market, ranging from non-profits and open-source communities to large private corporations and governmental institutions, each of which come with varying roles in the AI system lifecycle (see, e.g., Barclay & Abramson, 2021). For the purposes of this paper, it suffices to establish that all organisations – whether private, public or non-profit – that provide or deploy AI systems are here considered to be subject to the principles of justice based on a) their profound impact on people's lives, b) the way in which they limit people's ability to act as free and equal beings, and c) their key role in securing just background conditions for people and associations to function in the society.

Characteristic of a contractarian, Rawls establishes an idea of a hypothetical *original position* to define the fairest conditions possible where the principles of justice would be agreed upon. In Rawls's original position, the principles of justice are defined by people behind a *veil of ignorance*: they know neither their position in the society, nor their personal attributes that might affect the distribution of rights and duties (Rawls, 1999, p. 11). Even one's conception of good is hidden, leaving the parties with knowledge only of factors that are relevant for justice (Rawls, 1999, p. 11). The parties in the original position are "rational and mutually disinterested" (Rawls, 1999, p. 12), meaning that they do not pay attention to other parties' interests or goals, but merely "prefer more primary social goods rather than less" for themselves, thus knowing that factors such as liberties, opportunities and means to promote their aims are worth defending (Rawls, 1999, p. 123). The parties are also equal in their conception of good and capability of sense of justice (Rawls, 1999, p. 17).

Consequently, the principles of justice defined by Rawls are justified by his arguments on the fairness of the original position. In this paper, I will not start drafting the ethics guidelines by returning to the original position but start by applying the principles defined by Rawls. However, Rawls meant the original position to be such that one can always enter it again (Rawls, 1999, p. 17) and by *reflective equilibrium* adjust the contractarian situation, as well as the principles resulting from it, to eventually end up with principles that are just in a particular context (Rawls, 1999, p. 18). In the context of AI and fairness, this adjusting process has been shown to be particularly relevant (see, e.g., Franke, 2021), which also shows in the analysis below.

When Rawls first published his theory of justice, AI ethics as a field of study was in its infancy with only a couple of notable exceptions (see e.g., Wiener, 1954; Weizenbaum, 1976; see also Bynum, 2006). Hence, Rawls himself has given little indications of the applicability of his theory to the context of AI. His theory has, however, been shown to serve as a viable alternative to utilitarian methods of embedding ethics into algorithms and robotics by, e,g, Leben (2017, 2018). Taking a Rawlsian approach can also be justified by the relevance of the concept of justice

in the current AI principles and guidelines (see Jobin et al., 2019). Several scholars also address justice related to biased algorithms and discrimination (Bigman et al., 2023; Hacker, 2018; Janssen et al., 2022; König & Wenzelburger, 2020; Tolan, 2019), leading to an emergence of a field called *algorithmic fairness* (Lepri et al., 2018; Mitchell et al., 2021). Therefore, it seems that justice and fairness are values that people in modern democracies strive for AI technologies being no exception—which makes Rawls's theory of justice a theoretically interesting starting point for ethical AI principles.

Rawls's theory has deservedly received critique. For instance, Sen discusses Rawls's perception of justice as fairness in his *Idea of Justice* (Sen, 2010) and argues that the theory is weakly applicable to global context, where the institutions are not national or international, but supranational and global (Anderson, 2003; Harsanyi, 1975). In addition, Harsanyi (1975) has voiced critique on Rawls's decision to use the maximin principle as a decision theory that guides the reasoning in the original position, and thus the formation of the principles of justice. Robert Nozick (2013 [1974]), on the other hand, responded to Rawls's theory of justice with an *entitlement theory* of justice. According to Nozick, instead of aiming for an equal distribution of basic goods, everyone should get what they are entitled to, whether through justice in acquisition, transfer or rectification of justice (Nozick, 2013, p. 150–153), thus proposing a libertarian view to distributive justice. In the context of AI and algorithmic systems, embedding Rawlsian ethics into algorithms has also been challenged (e.g., critique on Leben's work in Keeling, 2018). While several of these limitations are discussed below, a further discussion on the moral theoretical soundness of Rawls's theory is a topic for another paper. These considerations in mind, we do not suggest this work to be the ultimate solution for fair development and deployment of AI systems but rather one contribution to the broader set of works striving towards ethical digital societies. In this discussion, Rawls still offers a useful starting point for reflecting the ethical foundations of AI in the context of democratic societies, which I demonstrate below.

Rawlsian Ethics Guidelines for Fair AI

In what follows, I apply Rawls's theory to form a set of ethics guidelines for fair AI. In the Rawlsian spirit, the guidelines here are destined to the basic structure of society (see Sect. 2) in which organisations develop and/or use AI systems.

Principle 1: Basic Liberties

According to the first principle of justice, each person should have equal right to "the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others" (Rawls, 1999, p. 53). Rawls offers a preliminary list of basic liberties in *A Theory of Justice* and further develops and justifies them in *Political Liberalism*. Accordingly, basic liberties to be equally distributed are:

- Freedom of thought and liberty of conscience
- Political liberties and freedom of association (the right to vote and to hold public office)
- Liberty and integrity of the person (including freedom from psychological oppression and physical assault and dismemberment)
- Liberties covered by the rule of law. (Rawls, 1999, p. 53; 2005, p. 335)

Rawls set his principles in an order of priority, meaning that neglecting any of the principles shall never be justified by fulfilling a subsequent principle. Applied to the context of AI, let us formulate the first Rawlsian principle for fair AI with the highest priority as follows:

1. Developers and deployers of an AI system must ensure that the AI system does not threaten the basic liberties of any individual.

Drawing from liberties Rawls himself has listed (Rawls, 1999, p. 53; 2005, p. 335), the guideline is hereby extended with a set of requirements that aim to bring Rawls's basic liberties into the context of development and application of AI. The first one is:

1.1 AI systems should not endanger but support the freedom of thought and liberty of conscience

Today, AI systems could be used to erode such freedom through surveillance (see Nemitz, 2018; Manheim & Kaplan, 2019; Wirtz et al., 2019; Zuboff, 2019; Véliz, 2021; Coeckelbergh, 2022a; Saura et al., 2022). Recent discussion on technologies that enable mass surveillance[1] demonstrates the global scope of the threat—even if not necessarily always realised—and how easy it is to mask. The discussion further intensified after the French government passed a regulation allowing AI powered surveillance during Paris 2024 Olympics.[2] Furthermore, even the mere threat of surveillance or nudging can be seen as a threat to freedom of thought because of how it changes our behaviour to advance someone else's goals rather than our own (Coeckelbergh, 2022a). Thus, following Rawls's requirement to prioritise protection of freedom of thought and liberty of conscience, AI ethics guidelines should guide the adoption of AI technologies so that they do not enable mass surveillance, or other types of manipulation or nudging that reduce our autonomy and agency. Considering the priority of the

[1] For example, a recent European Digital Rights (EDRi) report revealed ongoing biometric mass surveillance in several European countries: https://edri.org/our-work/new-edri-report-reveals-depths-of-biometric-mass-surveillance-in-germany-the-netherlands-and-poland/. These technologies are increasingly using AI technologies to collect, process and redistribute data.

[2] A piece of news by Reuters (23 March 2023): https://www.reuters.com/technology/france-looks-ai-powered-surveillance-secure-olympics-2023-03-23/.

first principle of justice and thus the basic liberties, this should not be done even in the benefit of the least advantaged members of the society.[3]

1.2 AI systems should not compromise but support political liberties and freedom of association, such as the right to vote and to hold public office

This liberty appears problematic since it is tied to democratic processes traditionally defined in national constitutions. AI technologies, on the other hand, are not always limited to national contexts. For example, generative AI systems like chatbots based on Large Language Models (LLMs), as well as AI-enhanced platforms, such as social media, reach a global audience. Ensuring the preservation of political liberties and freedom of association may be easier for national governmental institutions that use AI tools only in national government. For any global organisation, on the other hand, further operationalisation is needed. As such, this is a discussion that would merit its own paper, which is why, in the scope of this paper, I only expand on few key issues of applicability.

As Sen (2010) points out as one of the major difficulties in Rawls's theory, it is poorly applicable to global contexts and global justice. Although Rawls does introduce a continuation to his theory of justice in *The Law of Peoples* that aims to address the need for international collaboration, it only extends to the context of political liberalism, offering "ideals and principles of the *foreign policy* of a reasonably just *liberal* people" (Rawls, 2001, p. 10) (emphasis by the original author). Its main agents are still peoples, each with their "own internal governments" (Rawls, 2001, p. 3). AI systems are not, however, on the market only for just and liberal people to use. In this situation, one could be drawn towards Rawls's *historical method* (Rawls, 2005, p. 292–293), to refine the liberties: one could analyse a set of democratic constitutions that exist in the target market of a particular AI technology and seek for best practices that are the most efficient in supporting political liberties and freedom of association. Yet, this would not address the issue that AI systems are not under control by the nation-states—liberal or not—but mostly private companies and/or non-/capped-profits that can operate under several different jurisdictions. Many of these jurisdictions belong to non-democratic—in Rawls's terms non-decent—governments.

Making sure that AI systems do not compromise but support political liberties and freedom of association is thus challenging and would require extensive international coordination. In Rawls's view, this would require that the basic institutions of non-liberal societies "meet certain specified conditions of political right and justice and lead its people to honor a reasonable and just law for the Society of Peoples" (Rawls, 2001, p. 60). To do so—assuming optimistically, if

[3] For example, a recent discussion on Apple's intention to install anti-child porn technology in iPhones has been widely considered a violation of fundamental rights. See, e.g., https://www.pcmag.com/news/report-apple-to-scan-iphones-for-child-sexual-abuse-imagery. It should, therefore, be primarily explored whether there are other measures to protect the integrity of the child (see liberty 1.3) other than digital mass surveillance, so that we would be most likely to follow the "preferred path to the social state in which all the basic liberties can be fully instituted" (Rawls, 1999, p. 132).

not naively, that it would be possible in the first place—would require unforeseen efforts in global politics.

Moreover, freedom of elections is challenged when AI is used to influence potential voters. Candidates can be promoted by interactive social bots that spread messages and communicate with other users in a way that strongly resembles human activity. These bots can be automated either by political campaigners or external sources and can lead to a candidate reaching an unproportionate advantage compared to others that do not have access to these technologies or consider them unacceptable because of their manipulatory nature (see e.g., Brkan, 2019; Coeckelbergh, 2022a, p. 28; see also Kilovaty, 2019; Manheim & Kaplan, 2019). Recent developments seem to have made possible significant increase in efficiency, interactivity, and thus manipulatory potential of such attempts. For instance, popularisation of generative AI technologies has made it considerably easier to produce deepfakes and misleading information, which is profoundly changing digital communication as a basis for political participation (see, e.g., Jones, 2023). This unfortunate potential has been already utilised in parliamentary elections in Europe (e.g., Maker, 2023; Spring, 2024). In Rawlsian terms, the use of AI for such purposes is in principle considered a violation of basic liberties, and hence, unfair. On the other hand, political participation can potentially be enhanced by developing AI tools that facilitate democratic deliberation and encourage political participation (Alnemr, 2020; Bernholz et al., 2021; Helbing, 2021). These elements could thus be highlighted to strengthen this element of basic liberties.

1.3 AI systems should not harm but support the liberty and integrity of the person, including freedom from psychological oppression and physical assault and dismemberment

In the light of recent research (see e.g.,Manheim & Kaplan, 2019; Kilovaty, 2019; König & Wenzelburger, 2020; Laitinen & Sahlgren, 2021; Prunkl, 2022), some of the AI technologies currently in use, such as algorithms for online manipulation and effective deep-fake technologies, are threatening the very autonomy of humans. For example, Kilovaty (2019, p. 469) considers online manipulation to "effectively deprive individuals of their agency by distorting and perverting the way in which individuals typically make decisions." Similarly, König and Wenzelburger (2020, p. 7) consider this as one of the most serious negative consequences that AI might impose on democracy since freedom and human autonomy are among the core democratic values. Furthermore, Coeckelbergh (2022a, p. 118–122) discusses how AI can be seen through the concept of performance, affecting and conducting our bodies as well as minds. Lastly, AI has already being used in military to automate warfare, which hold several ethical issues with potential to physical oppression and assault (Forrest et al., 2020). For Rawls, such AI systems would be in principle deemed unethical, at least to an extent to which they take us further away from the goal of reaching equal basic liberties for everyone rather than advance it. As shown by, e.g., Johansson (2018),

use of autonomous weapons could reduce casualties and thus work in favour of 1.3. in some cases. We have thus once again arrived in a situation where further consideration is needed to contextualise the principle in individual use cases.

1.4 All AI systems should be aligned with the principle of rule of law

Accordingly, AI systems should be used to support, for example, due process and fundamental rights. Current tools adopted by certain judiciary institutions, such as COMPAS used in the US, seem to violate this principle by treating people differently based on, e.g., their ethnicity (on ethical issues concerning AI in jurisprudence, see e.g., Larson et al., 2016). Similarly, the use of AI systems has already led to unlawful arrests of people (see, e.g., Coeckelbergh, 2024, p. 42). These are examples of AI tools that violate the basic liberty of rule of law instead of supporting it, which demonstrates that more consideration is needed in institutions with a key role in ensuring the rule of law when applying AI systems in order for them to align their actions with Rawlsian principles of justice.

One quickly observes that Rawls's liberties are not free of conflicts. Let us consider, for instance, the persisting conflict between principles 1.1 and 1.3: exercising the freedom from physical assault would benefit from continuous AI-powered mass surveillance to target, e.g., potential terrorist attacks or rapes. This would, however, violate the freedom of thought and liberty of consciousness. The same applies for AI used in military, where the potential harm to the rule of law (1.4) could be estimated minimal compared to the potential positive impact on the liberty and integrity of the person (1.3). Rawls takes note of this issue by stating that basic liberties can be limited "when social circumstances do not allow the effective establishment of these basic liberties", but only with the one condition: "these restrictions can be granted only to the extent that they are necessary to prepare the way for the time when they are no longer justified" (Rawls, 1999, p. 132). He continues by suggesting that even if such compromises are necessary, the circumstances should still be "sufficiently favorable so that the priority of the first principle points out the most urgent changes and identifies the preferred path to the social state in which all the basic liberties can be fully instituted" (Rawls, 1999, p. 132).

In other words, Rawls's list of liberties is not absolute – they can and maybe even need to be modified to fit the "social, economic and technological" context of the society under scrutiny (Rawls, 1999, p. 54). He offers, as a response to critique, two methods for defining an appropriate set of basic liberties: 1) a *historical* survey of democratic constitutions to see which ones have traditionally worked well, and 2) an *analytical* consideration based on exhaustion on the set of liberties that are essential for principles agreed upon in the original position (Rawls, 2005, p. 292–293). The second method means developing repeatedly iterations of the list of basic liberties, eventually ending up with one that is fair according to the criteria of the original position. These methods and possible others have been discussed by several academics over the years (see e.g., McLeod & Tanyi, 2021), and could be used to set more refined and contextualised guidelines on the basis of the principles discussed in this paper. Here, the

guidelines are drafted with the freedoms listed by Rawls, leaving their further development to future research.

The Second Principle: Equality of Opportunity and Difference Principle

The second principle is divided into two sub-principles that apply to the "distribution of income and wealth and to the design of organizations that make use of differences in authority and responsibility" (Rawls, 1999, p. 53). First, all social and economic inequalities "must be attached to offices and positions open to all under conditions of fair equality of opportunity" (*equal opportunity principle*) (Rawls, 2005, p. 291). Hence, the basic structure of society must guarantee that each individual has equal opportunities to access the advantageous positions. This is particularly relevant in the context of choices regarding recruitment and performance evaluation, as well as access to education. However, events that influence people's paths to become eligible for a certain position in the first place also needs to be given attention. For example, a simple recommendation algorithm could influence the choices of people when it comes to skill development or decisions to seek for positions in certain fields (e.g., Kokkodis & Ipeirotis, 2021). When biased, such algorithms could direct some people towards less advantageous positions than others on an unfair basis. This could also lead discrimination of people working certain professions over others – the future of jobs such as freelancer designers and writers has already raised questions of impacts on individuals' livelihoods and opportunities in life (e.g., Mims, 2024).

Rawlsian guidelines could thus recommend the following:

2. The use and development of AI systems should not negatively impact people's opportunities to seek income and wealth. If an AI system is used in distribution of advantageous positions, such as recruitment, performance evaluation, or access to education, it needs to be ensured that.

    2.1 the tool is trained with non-biased training data, or appropriate tools are used to mitigate the biases in the final product if no non-biased training data is available (data bias mitigation),
    2.2 the outcome of the use of the tool includes an explanation of the grounds for the outcome it produces (explainability), and.
    2.3 the algorithms used shall encourage neither biased results nor the systematic repetition and amplification thereof in, e.g., the feedback loops of a machine learning system (algorithmic bias mitigation).

If these conditions cannot be met, AI should not be used in the process.

In the light of datasets currently used in such processes, these requirements are likely to be close to impossible to meet. As Russell (2019, p. 129) explains, the bias is often rooted in the culture where the data has been created and extracted. Simply removing the factors that are not supposed to affect the decision—such

as race, age, or gender—is not sufficient because of redundant encodings that enable biased prediction from other aspects (Hardt et al., 2016). Although several suggestions have been made about fairness evaluation and correction models for AI datasets (see e.g., Hardt et al., 2016; Mehrabi et al., 2022), there is no evidence implying they were widely adopted. Even if the lack of transparency is not an issue per se and can be challenged as an indispensable value (see, e.g., Holm, 2019), in the context of Rawlsian equality of opportunity, a meaningful level of explainability would be required to avoid systematic biases that constitute a persisting problem in AI-assisted recruitment (e.g., Kazim et al., 2021; Tilmes, 2022).

Yet, one could argue that human decisions are not free of bias, either, and thus using an AI system could lead to decisions that support the equality of opportunity compared to human decision-making (see, e.g., Zerilli et al., 2019). However, from a Rawlsian perspective, the human decision-making should also aim at equality of opportunity, and thus the failure of, e.g., a recruiter to follow the principle should not justify any lesser principle for the use of AI systems. Moreover, AI systems tend to systematise biases towards certain groups of people, which means that the potential to improve the equality of opportunity of some individuals might happen to the detriment of those in particularly vulnerable positions. As we will see below, this can pose a problem also in the light of Rawls's difference principle. To add to the complexity, for the basic structure of society to ensure people do not lose their livelihoods due to AI would also require extensive measures in, e.g., policy, to ensure changes in labor markets would be appropriately managed and meaningful opportunities found to replace the lost jobs. Therefore, defining the responsibilities between institutions belonging to the basic structure would require more attention, which will be further discussed below.

Characteristic to Rawls's theory, this principle aims at an ideal situation rather than incremental change, which has deservedly received critique (e.g., Sen, 2010). It also neglects all interpretations of fairness other than Rawls's, which are frequently discussed amongst researchers of algorithmic fairness (see, e.g., Chouldechova, 2017; Baumann & Loi, 2023; Green, 2022). Consequently, considering the above-discussed need to adjust the list of basic liberties according to their context, deciding upon the biases that should in this ideal situation be mitigated is most likely not universal. Still, as for our experiment following Rawls, we rely on his conception of fairness and maintain that if the use of AI indicates a risk of systematic erosion of equality of opportunity for some individuals or groups thereof, organisations should not use such AI systems in recruitment or other distribution of advantageous positions, no matter the potential increase in efficiency that the tool could bring.

As for the latter part of the second principle, *difference principle*, inequalities "must be to the greatest benefit of the least advantaged members of society" (Rawls, 2005, p. 291). This principle applies to the "distribution of income and wealth and to the design of organizations that make use of differences in authority and responsibility" (Rawls, 1999, p. 53). To apply this principle to the context of AI, the following recommendation is given:

3. All inequalities affected by AI systems, such as acquiring a position of power or accumulation of wealth, must be to the greatest benefit of the least advantaged members of society.

The development of AI technologies has led to concentration of power over many aspects of our lives to tech giants (Coeckelbergh, 2022a; Nemitz, 2018, p. 2–3), which means that the interests of the shareholders of these companies have an increasing influence on the distribution of inequalities, as well as on algorithms that are used to determine the value basis for doing so. Rawls's difference principle does not merely prohibit advantages detrimental to the least advantaged but requires an *increase* in benefit for the least advantaged as a result of any such inequality, which reflects the maximin principle Rawls uses as the decision theory on choosing one's action in an event of uncertainty. Accordingly, we should make decisions about the direction of AI development based on an evaluation of the worst possible scenario of each option, and choose an action the worst scenario of which is the least bad. Although Rawls's use of the maximin principle has been heavily and perhaps deservedly criticised by, e.g., Harsanyi (1975) and further developed by Sen (2010), Rawls persistently defended his version, leading to a continuous discussion notably in the field of economics about its potential to form a basis for fair distribution of resources (e.g., Mongin & Pivato, 2021). Therefore, without going into details of these arguments, Rawls's perspective is here applied as is.

The current accumulation of wealth from AI that has shown often to harm rather than enhance the lives of the least advantaged (e.g., discriminatory systems used to make life-changing decisions in Tilmes, 2022; AI systems that undermine minorities in, e.g., Bender et al., 2021; Janssen et al., 2022; or AI systems eroding societal structures that protect fundamental rights in Coeckelbergh, 2024), which seems contradictory to the difference principle. However, many of these systems would already be deemed unfair on the basis of principles 1 or 2. Having in mind that Rawls's principles come in an order of priority, for a system to be considered against the difference principle, it should first be aligned with the preceding principles. For example, an AI system deployed by a financial institution could be used to determine the probability of a prospective lender to fail the contract terms. Even if the system was built in full alignment with the basic liberties and the biases were mitigated appropriately, if the system benefits most the shareholders of the banking institution (and does not, for example, improve the position of the least advantaged in mitigating human biases in loan decisions), it fails in regard to the difference principle.

It is obvious that showing or estimating the implications for the least advantaged in the society in a precise manner is complicated, if not impossible in modern societies. It would first require defining who are the least advantaged, which already poses a great difficulty. To align AI with this principle would thus require active reflection on the impacts of AI systems on different groups of people in both development and deployment. For instance, a company developing a generative, general purpose AI system would need to conduct broad impact analyses and reflect the consequences

of their actions as well as their moral duties in a broader scope, which brings them closer to a situation similar to deliberation in an original position.

Therefore, following Rawls's theory, AI systems should always thus encourage societal improvement when used in processes that lead to inequalities (as opposed to "levelling down" as a result of fairness algorithms, see Mittelstadt et al., 2023). These guidelines would thus require institutions to review their existing approaches, such as data-capitalist business logic for wealth maximization, to align their actions with the Rawlsian principles. Regulating institutions and policy-makers could seek for solutions from regulation that set corresponding minimum requirements for AI systems, as well as standards for enforcement—the soon-to-be-adopted EU AI Act being one (although non-Rawlsian) example of a supranational approach. Private corporations could include the difference principle into their frameworks for innovation so that no system that fails to comply gets to production.

Even then, we would be left with fundamental questions: What kind of balance between state regulation and private power leads to the fairest AI systems overall? How do the measures by private and public institutions tie together into a functional societal order? Will that system still resemble a democracy, which in Rawls's perspective is the only legitimate form of fair societies? As Coeckelbergh (2024, p. 76) notes, it is not only a question of how much regulation we need but what kind of institutional structure we have in place to implement the policies and principles, and how we change our technologies themselves. Could Coeckelbergh's technodemocracy (2024, p. 83–83) offer us the answers we are looking for? How about Susskind's (2022) digital republic or Muldoon's (2022) platform socialism? Answering these questions would require broader interdisciplinary discussion and policy research that exceed the scope of this paper. As for now, these guidelines offer a reference point for the institutions in the basic structure of society to evaluate the fairness of their development and deployment of AI systems.

Lastly, it is worth noting that due to the context where Rawls created his principles, these guidelines are drafted with liberal democracies in mind, which means that there's a strong caveat in their universal applicability. This is characteristic to Rawls's theory of justice and becomes especially apparent in his later developments in *The Law of Peoples* (Rawls, 2001), which has received critique by, e.g., Sen (2010). Now, it has been further illustrated in the application of Rawls's principles to the context of AI, which indicates a need for further research in ethics perspectives that would better grasp the global nature of AI and thus the need for ethical principles that consider the societal structures in regimes other than democracies, as well.

Conclusions and Discussion

The goal of this paper has been to experiment with a set of ethics guidelines for AI following John Rawls's theory of justice as fairness. Today, major ethical concerns of AI technologies are related to inequalities and other issues that threaten the future of democratic governance (see e.g., Brkan, 2019; Coeckelbergh, 2022a, 2022b, 2024; König & Wenzelburger, 2020). Meanwhile, the major AI ethics guidelines

introduced so far have shown to neglect the relationship between AI and democracy (see e.g., Hagendorff, 2020) and to lack in disclosing the justifications – the very ethics – behind the principles (Franzke, 2022; Hagendorff, 2020; Jobin et al., 2019). The current paper is an attempt towards strengthening the moral-philosophical basis of the principles that serve as the foundation for applications and operationalisation of ethics amongst both the providers and users of AI systems. This is an exploration of the possibility of creating an ethics guideline that aims at ensuring societal stability and societal structures that support the ethicality of AI. Rawls being one of the most important modern philosophers with the goal of establishing principles of justice for modern democracies, his theory provides a fruitful basis for this exploration. The conclusive set of Rawlsian ethics guidelines for fair AI are:

1. Developers and deployers of an AI system must ensure that the AI system does not threaten the basic liberties of any individual.

    1.1 AI systems should not endanger but support the freedom of thought and liberty of conscience.
    1.2 AI systems should not compromise but support political liberties and freedom of association, such as the right to vote and to hold public office.
    1.3 AI systems should not harm but support the liberty and integrity of the person, including freedom from psychological oppression and physical assault and dismemberment.
    1.4 All AI systems should be aligned with the principle of rule of law.

2. The use and development of AI systems should not negatively impact people's opportunities to seek income and wealth. If an AI system is used in distribution of advantageous positions, such as recruitment, performance evaluation, or access to education, it needs to be ensured that.

    2.1 The tool is trained with non-biased training data, or appropriate tools are used to mitigate the biases in the final product if no non-biased training data is available (data bias mitigation),
    2.2 The outcome of the use of the tool includes an explanation of the grounds for the outcome it produces (explainability), and.
    2.3 the algorithms used shall encourage neither biased results nor the systematic repetition and amplification thereof in, e.g., the feedback loops of a machine learning system (algorithmic bias mitigation).

If these conditions cannot be met, AI should not be used in the process.

3. All inequalities affected by AI systems, such as acquiring a position of power or accumulation of wealth, must be to the greatest benefit of the least advantaged members of society.

Are these guidelines any better than the ones already drafted? In choosing Rawls's theory as a basis for the guidelines, I have addressed several issues

found in the existing guidelines, the first being previously mentioned as the lack of discussing democracy. These guidelines are also improved in their ability to prioritise: whereas existing guidelines do not offer solutions in cases of conflicting principles (Jobin et al., 2019), Rawlsian guidelines are hierarchical, which helps with resolving priority conflicts. Yet, the analysis shows that a number of conflicts still persist—the Rawlsian approach is not an off-the-shelf solution, or a silver bullet that would eradicate the need for contextualising the principles and exercising active ethical reflection while developing and deploying AI systems, which is, as e.g. Rességuier and Rodrigues (2020) and Heilinger (2022) note, is an essential part of effective application of ethics in the context of AI. To do so, approaches such as those of Rahwan (2018) that highlight the need for public engagement and deliberation could ensure that the principles rooted in rigorous ethics get operationalised in an actionable way, which is a topic for future research.

Furthermore, Jobin et al. (2019) point out that the existing guidelines focused on negative characterisation of ethical values, and thus tended to neglect the possibility of promoting favorable values with AI, not just ensuring the principles are met *despite* of the use of AI in the context in question. In these guidelines, this is considered in guidelines 1.1, 1.2 and 1.3 as well as in the guideline 3, embedded in the difference principle. Taking Rawls's theory, which is rooted in freedom and equality, as a starting point also addresses what Jobin et al. (2019) identified as another flaw in existing guidelines: they lack in discussing human dignity and solidarity. These two topics are embedded in Rawlsian principles—freedom and equality are fundamental aspects that aim at guaranteeing equal basic liberties for all, and the difference principle guides towards consideration of the least advantaged members of society besides one's own individual interests. The difference principle also addresses the frequent issue with algorithmic fairness pointed out by Mittelstadt et al. (2023) in their pre-print, according to which strictly egalitarian views of fairness easily lead to "levelling down", i.e., making everyone worse off. Rawlsian guidelines also offer a philosophically justified ground called for by Franzke (2022), and keep in mind the societal context created by democratic institutions—the lack of which was previously noted by Hagendorff (2020).

What these Rawlsian guidelines are arguably lacking are practical examples and applications in technology, as well as metrics to follow how well the principles are being fulfilled. They are in a high level of abstraction, meaning institutions need to reflect on what they mean in practice in their sectorial context. It must also be noted that these guidelines aim at applying and adapting Rawls's principles to today's society that differs in certain essential points from that of the time Rawls initially started working on his theory of justice, which stresses the importance of subjecting the principles to broader public discussion. That is, however, not an issue per se, as drafting principles is only the first part of the process (see e.g., Mittelstadt, 2019; Rességuier & Rodrigues, 2020), and there is currently an active academic discussion around operationalization of AI ethics principles and recommendations (see, e.g., Ayling & Chapman, 2022; Ibáñez & Olmeda, 2022; Bleher & Braun, 2023). The next step would be to subject the guidelines to public deliberation, bring them into practice and test them in real-life conditions, such as those described by Leben (e.g.,

2017, 2018) in his work on application of Rawlsian ethics to autonomous vehicles. The principles could be applied by design scientists to find ethically and societally sustainable practices for future AI development, and by practitioners to innovate use cases that are ethically sustainable. The principles could also be tested in efforts to create auditing frameworks for technologies and processes utilising AI technologies, as well as opening a broader dialogue on societal structures that would allow AI development to benefit most the least advantaged. Attaching a list of best practices and further developments would add to the applicability of these guidelines in practice.

One could also question whether subjecting all nations to a Western liberalist normative framework would be desirable or ethical in the first place, as digitalisation has already raised concerns about new forms of colonialist practices (e.g., Adams, 2021; Arora et al., 2023; Couldry & Meijas, 2019; Westerstrand et al., 2024). Rawls's premise that justice can only be achieved by democratic governments is reflected in the principles (e.g., 1.2.), which would require non-democratic governments to first adopt a new political system in order for them (or the companies operating under their legislature) to produce fair AI systems. This would be unreasonable. It is thus important to see the principles in their normative context and critically discuss their applicability in a global level. Consequently, these guidelines still remain—as is to be expected—at the principle-formation-level, and thus are not alone a solution for ethical AI.

Yet, I hope that these guidelines inspire both academics and practitioners to revisit their own existing guidelines and to evaluate how they relate to the Rawlsian principles. Did they differ in emphases, did they add or omit something? Is there a justification for inclusion or omission of some principles, or values? How about the larger societal context, would your organisation's guidelines support preservation of societal structures that promote fairness? Even though we need to keep bringing principles into practice, I argue that re-evaluating the moral-philosophical grounds of the existing principles and guidelines is still a vital part of practicing ethics in the dynamic, ever-changing context of AI development.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

References

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews, 46*(1–2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Albert, E. T. (2019). AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review, 18*(5), 215–221. https://doi.org/10.1108/SHR-04-2019-0024

Alnemr, N. (2020). Emancipation cannot be programmed: Blind spots of algorithmic facilitation in online deliberation. *Contemporary Politics, 26*(5), 531–552. https://doi.org/10.1080/13569775.2020.1791306

Anderson, B. C. (2003). The antipolitical philosophy of John Rawls. *The Public Interest, 151*, 39–51.

Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization, 33*(3), 100478. https://doi.org/10.1016/j.infoandorg.2023.100478

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management, 62*, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics, 2*(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Barclay, I., & Abramson, W. (2021). Identifying roles, requirements and responsibilities in trustworthy AI systems. In UbiComp/ISWC '21 Adjunct: Adjunct proceedings of the 2021 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2021 ACM international symposium on wearable computers (pp. 264–271). https://doi.org/10.1145/3460418.3479344.

Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, *36*, 45. https://doi.org/10.1007/s13347-023-00624-9

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). https://doi.org/10.1145/3442188.3445922.

Berkey, B. (2021). Rawlsian institutionalism and business ethics: Does it matter whether corporations are part of the basic structure of society? *Business Ethics Quarterly, 31*(2), 179–209. https://doi.org/10.1017/beq.2020.14

Bernholz, L., Landemore, H., & Reich, R. (2021). *Digital technology and democratic theory*. University of Chicago Press.

Bigman, Y. E., Wilson, D., Arnestad, M., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General, 152*(1), 4–27.

Blanc, S., & Al-Amoudi, I. (2013). Corporate institutions in a weakened welfare state: A Rawlsian perspective. *Business Ethics Quarterly, 23*(4), 497–525. https://doi.org/10.5840/beq201323438

Bleher, H., & Braun, M. (2023). Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics, 29*, 21. https://doi.org/10.1007/s11948-023-00443-3

Bostrom, N. (2016). The control problem. Excerpts from *Superintelligence: Paths, dangers, strategies*. In S. Schneider (Ed.), *Science fiction and philosophy* (pp. 308–330). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118922590.ch23

Bostrom, N. (2017). *Superintelligence: Paths, dangers*. Dunod.

Brkan, M. (2019). Artificial intelligence and democracy: *Delphi—Interdisciplinary Review of Emerging Technologies*, *2*(2), 66–71. https://doi.org/10.21552/delphi/2019/2/4

Bynum, T. W. (2006). Flourishing ethics. *Ethics and Information Technology, 8*(4), 157–173. https://doi.org/10.1007/s10676-006-9107-1

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data,* 5(2). https://doi.org/10.1089/big.2016.0047.

Clouser, K. D., & Gert, B. (1990). A critique of principlism. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 15*(2), 219–236. https://doi.org/10.1093/jmp/15.2.219

Coeckelbergh, M. (2022a). *The political philosophy of AI: An introduction*. John Wiley & Sons.

Coeckelbergh, M. (2022b). Democracy, epistemic agency, and AI: Political epistemology in times of artificial intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00239-4

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

Couldry, N., & Meijas, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television and New Media, 20*(4), 336–349.

Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior, 116*, 106626. https://doi.org/10.1016/j.chb.2020.106626

Forrest, M., Boudreaux, B., Lohn, A., Ashby, M., Christian, C., & Klima, K. (2020). *Military applications of artificial intelligence: Ethical concerns in an uncertain world*. https://apps.dtic.mil/sti/citations/AD1097313

Franke, U. (2021). Rawls's original position and algorithmic fairness. *Philosophy and Technology, 34*(4), 1803–1817. https://doi.org/10.1007/s13347-021-00488-x

Franzke, A. S. (2022). An exploratory qualitative analysis of AI ethics guidelines. *Journal of Information, Communication and Ethics in Society, 20*(4), 401–423. https://doi.org/10.1108/JICES-12-2020-0125

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Green, B. (2022). Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy Technology*, *35*, 90. https://doi.org/10.1007/s13347-022-00584-6

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945. https://doi.org/10.2307/20439056

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(4). https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\COLA\COLA2018095.pdf

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review, 69*(2), 594–606. https://doi.org/10.2307/1959090

Heidari, H., Loi, M., Gummadi, K. P. & Krause, A. (2019). A moral framework for undestanding fair ML through economic models of equality of opportunity. In Proceedings of the conference on fairness, accountability, and transparency (FAT* '19) (pp. 181-190). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3287560.3287584

Heilinger, J.-C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy And Technology, 35*(3), 61. https://doi.org/10.1007/s13347-022-00557-9

Helbing, D. (2021). *Next civilization: Digital democracy and socio-ecological finance-how to avoid dystopia and upgrade society by digital means*. Springer.

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics, 1*(1), 41–47. https://doi.org/10.1007/s43681-020-00008-1

Holm, E. (2019). In defense of the black box. *Science, 364*(6435), 26–27.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics, 178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY, 37*(4), 1663–1687. https://doi.org/10.1007/s00146-021-01267-0

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will algorithms blind people? The effect of explainable ai and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review, 40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2.

Johansson, L. (2018). Ethical aspscts of military maritime and aerial autonomous systems. *Journal of Military Ethics, 17*(2–3), 140–155. https://doi.org/10.1080/15027570.2018.1552512

Jones, N. (2023). How to stop AI deepfakes from sinking society—and science. *Nature*, *621*, 676–679. Available at: https://doi-org.ezproxy.utu.fi/https://doi.org/10.1038/d41586-023-02990-y

Kazim, E., Koshiyama, A. S., Hilliard, A., & Polle, R. (2021). Systematizing audit in algorithmic recruitment. *Journal of Intelligence, 9*(3), 46.

Keeling, G. (2018). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In Müller, V. (eds) *Philosophy and theory of artificial intelligence 2017 (PT-AI 2017). Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol 44. Springer. https://doi.org/10.1007/978-3-319-96448-5_29

Kilovaty, I. (2019). Legally cognizable manipulation. *Berkeley Technology Law Journal*, *34*, 449.

König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly, 37*(3), 101489. https://doi.org/10.1016/j.giq.2020.101489

Kokkodis, M., & Ipeirotis, P. G. (2021). Demand-aware career path recommendations: A reinforcement learning approach. *Management Science, 67*(7), 4362–4383. https://doi.org/10.1287/mnsc.2020.3727

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Laitinen, A., & Sahlgren, O. (2021). AI systems and respect for human autonomy. *Frontiers in Artificial Intelligence, 4*(705164), 1–14. https://doi.org/10.3389/frai.2021.705164

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1), 3–3.

Leben, D. (2017). A Rawlsian apgorithm for autonomous vehicles. *Ethics of Information Technology, 19*, 107–115. https://doi.org/10.1007/s10676-017-9419-3

Leben, D. (2018). *Ethics for robots*. Routledge.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Maker, M. (3 October 2023). Slovakia's election deepfakes show AI is a danger to democracy. *Wired*. https://www.wired.co.uk/article/slovakia-election-deepfakes. Last accessed on 10 July 2024.

Manheim, K., & Kaplan, L. (2019). *Artificial intelligence: Risks to privacy and democracy*. *21*.

McLeod, S. K., & Tanyi, A. (2021). The basic liberties: An essay on analytical specification. *European Journal of Political Theory*, 14748851211041702. https://doi.org/10.1177/14748851211041702

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Mims, C. (21 June 2024). AI doesn't kill jobs? Tell that to freelancers. *Wall Street Journal*, https://www.wsj.com/tech/ai/ai-replace-freelance-jobs-51807bc7. Last accessed 10 July 2024.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), Article 11. https://doi.org/10.1038/s42256-019-0114-4.

Mittelstadt, B., Wachter, S., Russell, C. (2023). The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. ArXiv pre-print. https://doi.org/10.48550/arXiv.2302.02404.

Mongin, P., & Pivato, M. (2021). Rawls's difference principle and maximin rule of allocation: A new analysis. *Economic Theory, 71*, 1499–1525. https://doi.org/10.1007/s00199-021-01344-x

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines, 31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society, 38*(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8

Muldoon, J. (2022). *Platform socialism*. Pluto Press.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180089. https://doi.org/10.1098/rsta.2018.0089

Nozick, R. (2013) [1974]. *Anarchy, state, and utopia.* Basic Books.

Paterson, T., & Hanley, L. (2020). Political warfare in the digital age: Cyber subversion, information operations and 'deep fakes.' *Australian Journal of International Affairs, 74*(4), 439–454. https://doi.org/10.1080/10357718.2020.1734772

Pitt, J. C. (2014). "Guns Don't Kill, People Kill"; Values in and/or around technologies. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 89–101). Springer. https://doi.org/10.1007/978-94-007-7914-3_6

Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence, 4*, 99–101. https://doi.org/10.1038/s42256-022-00449-9

Qamar, Y., Agrawal, R. K., Samad, T. A., & Chiappetta Jabbour, C. J. (2021). When technology meets people: The interplay of artificial intelligence and human resource management. *Journal of Enterprise Information Management, 34*(5), 1339–1370. https://doi.org/10.1108/JEIM-11-2020-0436

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*, 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rawls, J. (1971). *A theory of justice: original edition*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v

Rawls, J. (1999). *A theory of justice: Revised edition*. Harvard University Press.

Rawls, J. (2001). *The law of peoples: With "The Idea of Public Reason Revisited."* Harvard University Press.

Rawls, J. (2005). *Political liberalism*. Columbia University Press.

Rességuier, A., & Rodrigues, R. (2020). *AI ethics should not remain toothless!* A call to bring back the teeth of ethics. *Big Data & Society, 7*(2), 205395172094254. https://doi.org/10.1177/2053951720942541

Robinson, N., Hardy, A., & Ertan, A. (2021). *Estonia: A curious and cautious approach to artificial intelligence and national security* (SSRN Scholarly Paper No. 4105328). https://doi.org/10.2139/ssrn.4105328

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

Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2022). Assessing behavioral data science privacy issues in government artificial intelligence deployment. *Government Information Quarterly, 39*(4), 101679. https://doi.org/10.1016/j.giq.2022.101679

Sen, A. (2010). *The idea of justice*. https://www.penguin.co.uk/books/56627/the-idea-of-justice-by-amartya-sen/9780141037851

Singer, A. (2015). There is no Rawlsian theory of corporate governance. *Business Ethics Quarterly, 25*(1), 65–92. https://doi.org/10.1017/beq.2015.1

Spring, M. (8 June 2024). X takes action on deepfake network smearing UK politicians after BBC investigation. *BBC*. https://www.bbc.com/news/articles/cq55gd8559eo. Last accessed 7 July 2024.

Stahl, B. C. (2022). From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI Ethics, 2*(1), 65–77. https://doi.org/10.1007/s43681-021-00080-1

Stiglitz, J. E. (2013). *The price of inequality*. Norton. https://wwnorton.com/books/the-price-of-inequality/

Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics, 27*(1), 15. https://doi.org/10.1007/s11948-020-00277-3

Susskind, J. (2022). *The digital republic.* Bloomsbury.

Tilmes, N. (2022). Disability, fairness, and algorighmic bias in recruitment. *Ethics of Information Technology, 24*(2), 21.

Tolan, S. (2019). *Fair and unbiased algorithmic decision making: Current state and future challenges* (arXiv:1901.04730). arXiv. https://doi.org/10.48550/arXiv.1901.04730

Vakkuri, V., Jantunen, M., Halme, E., Kemell, K.-K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). *Time for AI (ethics) maturity model is now* (arXiv:2101.12701). arXiv. https://doi.org/10.48550/arXiv.2101.12701

Véliz, C. (2021). *Privacy is power.* Melville House.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co.

Westerstrand, S. (2023). Ethics in the intersection of AI and democracy: The AIDEM framework. *ECIS 2023 Research Papers*. https://aisel.aisnet.org/ecis2023_rp/321

Westerstrand, S., Westerstrand, R., & Koskinen, J. (2024). Talking existential risk into being: A Habermasian critical discourse perspective to AI hype. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00464-z

Wiener, N. (1954). *The human use of human beings: Cybernetics and society* ([2d ed. rev.]). Doubleday.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorighmic and human decision-making: Is there a double standard? *Philosophy and Technology, 32*, 661–683.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

Zuboff, S. (2019). *The age of surveillance capitalism*. https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/?lens=publicaffairs