AI and Ethics
https://doi.org/10.1007/s43681-023-00341-1

ORIGINAL RESEARCH

Participation, prediction, and publicity: avoiding the pitfalls of applying Rawlsian ethics to AI

Morten Bay[1]

Received: 27 May 2023 / Accepted: 23 August 2023

Abstract
Given the popularity of John Rawls' theory of justice as fairness as an ethical framework in the artificial intelligence (AI) field, this article examines how the theory fits with three different conceptual applications of AI technology. First, the article discusses a proposition by Ashrafian to let an AI agent perform the deliberation that produces a Rawlsian social contract governing humans. The discussion demonstrates the inviability of such an application as it contradicts foundational aspects of Rawls' theories. An exploration of more viable applications of Rawlsian theory in the AI context follows, introducing the distinction between *intrinsic* and *extrinsic* theoretical adherence, i.e., the difference between approaches integrating Rawlsian theory in the system design and those situating AI systems in Rawls-consistent policy/legislative frameworks. The article uses emerging AI legislation in the EU and the U.S. as well as Gabriel's argument for adopting Rawls' *publicity* criterion in the AI field as examples of extrinsic adherence to Rawlsian theory. A discussion of the epistemological challenges of predictive AI systems then illustrates some implications of intrinsic adherence to Rawlsian theory. While AI systems can make short-term predictions about human behavior with intrinsic adherence to Rawls' theory of justice as fairness, long-term, large-scale predictions results do not adhere to the theory, but instead constitute the type of utilitarianism Rawls vehemently opposed. The article concludes with an overview of the implications of these arguments for policymakers and regulators.

Keywords John Rawls · AI regulation · AI policy · Prediction · Democratic participation · Difference principle

1 Introduction

Since John Rawls is considered by many to be the most influential political and moral philosopher of the twentieth century, it is perhaps unsurprising that "researchers and industry developers in artificial intelligence (AI) and natural language processing (NLP) have uniformly adopted a Rawlsian definition of fairness" [1, p. 1]. Rawls' theory of justice as fairness is appealing due to the apparent simplicity of well-known concepts such as the *veil of ignorance* and the *original position*, and different scholars both promote the applicability of Rawlsian theory to AI ethics [2–4] and criticize it [1, 5]. However, scholars sometimes take Rawls' concepts out of their original context, dismissing intricate (but essential) theoretical components. The latter may even include conditions upon which the viability of the entire theory hinges according to Rawls himself. It seems that more clarity on how Rawlsian theory can apply to different instantiations of AI technology is needed, and this article aims to shed at least some light on this. First, I use a critique of an article by Ashrafian to exemplify and illustrate why it is inconsistent with Rawlsian principles to leave the deliberations involved in the construction of a Rawlsian social contract to an AI agent, regardless of how Rawlsian its algorithmic design is, which models it uses, or how these are trained. I then propose to classify integrations of Rawlsian principles and AI systems as either employing *extrinsic* or *intrinsic* adherence to Rawlsian theory, with the two terms, respectively, denoting whether the Rawlsian principles guiding the systems are part of the AI system's design or part of the policy/legislative framework in which it is situated (or both). To illustrate *extrinsic* adherence to Rawls' theory of justice as fairness, I analyze Gabriel's [6] explorations of the Rawlsian notions of *background institutions* and *publicity* in addition to analyzing emerging AI legislation in the EU and the U.S. Following this, I demonstrate how Rawls' arguments against utilitarianism (in combination with Karl Popper's

✉ Morten Bay
mortench@usc.edu

[1] Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, CA, USA

Published online: 05 September 2023

epistemological distinction between scientific prediction and *prophecy*) have implications for *intrinsic* adherence. That is, while short-term predictions about human behavior can be congruent with Rawlsian principles, AI systems performing large-scale predictions of human behavior over the long term violate core principles of Rawls' theory of justice as fairness.

2 AI-crafted social contracts and Rawls

As Powers and Ganascia [7] state, there are "conceptual ambiguities" [7, p. 29] in the language surrounding AI ethics. Emerging legislation such as the European Union's AI Act[1] attempts to reduce some of this ambiguity, making it an appropriate source of definitions. Per the European Commission, "AI" is currently used as a "'blanket term' for various computer applications…which exhibit capabilities commonly and currently associated with human intelligence" [8, p. 3]. Following this, my use of "AI" without additional qualifiers refers to this umbrella term and the field it covers. At the next level of specificity, I differentiate between *AI systems* and *AI agents.* The EU's AI Act[2] defines an *AI system* as"… software that is developed with [specific] techniques and approaches…and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" [8, p. 4]. In this article, AI systems generating all four types of outputs are examined with their individual contexts specified. I use the definition of an *AI agent* from Powers and Ganascia, denoting a computational entity, usually part of an AI system, that has *agency* but not *intention*. In contrast to the traditional Ethics view of an agent that "intends (upon reflection) its actions," the authors quote Russell and Norvig [9] and state that an AI "agent implements a function that maps percept sequences to actions." This difference between intentional actions and actions performed intentionlessly[3] by the actor, as prescribed by a third party, has "important consequences from an ethical point of view," according to Powers and Ganascia, as "an AI agent lacks true proper goals, personal intentions, or real freedom…" [7, p. 30]. The latter, as we shall see, is also important for the application of Rawlsian ethics to AI systems.

Ashrafian pursues such an application of Rawlsian theory in a well-intentioned and original proposition [10], which, as I will show in the following, is unfortunately also inconsistent with lesser-known, but essential aspects of Rawlsian theory. Ashrafian proposes an AI agent, designed to adhere to Rawls' ethics, that produces a social contract governing a society of humans. This objective, however, is undermined by strong conditions set by Rawls that Ashrafian appears to ignore. His proposition illuminates the difference between using well-known concepts such as the *original position* and the *veil of ignorance* as inspiration for creating new moral frameworks in the AI field (which has been done with some success in [2–4, 11, 12]) and attempting to apply Rawlsian theory comprehensively to a specific AI system. Unlike the former, the latter requires consideration of the conditions under which Rawls considers his theory valid.

Ashrafian's main argument revolves around the idea that an AI agent may be better suited than humans to construct a social contract for a "fair, just, and humane society" [10, p. 7] because its algorithms can be designed to be unconcerned with the interests that must otherwise be deliberately disregarded when establishing a social contract based on the Rawlsian principle of the *veil of ignorance*. Ashrafian thus applies the theory of justice as fairness at the scale Rawls intended, i.e., at the highest overall level of societal arrangements. It is important to note that Rawls never intended concepts such as the *veil of ignorance*, the *original position,* or his *Difference Principle*[4] to be used in low-level decision-making. He writes of the latter that it is a macro, not a micro principle" [13, p. 226]. Rather, Rawls insisted on limiting the ethical scope of his theory to overarching political questions concerning how a society can become *just* and *well-ordered* as a "fair system of cooperation" through a *basic structure* [14, p. 89]. He writes:

> "Justice as fairness hopes to extend the idea of a fair agreement to the basic structure itself. Here we face a serious difficulty for any political conception of justice that uses the idea of contract, whether or not the contract is social. The difficulty is this: we must specify a point of view from which a fair agreement between free and equal persons can be reached; but this point of view must be removed from and not distorted by the particular features and circumstances of the existing basic structure. The original position, with the feature I have called the "veil of ignorance"…specifies this point of view" [13, p. 15].

From this paragraph, some assertions can be drawn. First, a Rawlsian social contract is a *fair agreement* reached by *free and equal persons.* Second, a fair *basic structure* is the

[1] Draft current as of August 2023.

[2] Draft current as of August 2023.

[3] As opposed to "unintentionally", which has connotations of randomness.

[4] The *Difference Principle* states that "Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society" [13, pp. 42–43].

main intended outcome of such a social contract. Otherwise, Rawls would not be concerned with setting aside the "features and circumstances" of the "existing basic structure" and extending "the idea of a fair agreement to the basic structure itself." Third, the *original position* and *veil of ignorance* concepts are directly linked to, and occur in the context of, deliberations over the *basic structure* of society, which Rawls calls "…the primary subject of justice." He further describes the *basic structure* as "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation" [14, p. 6].

As a deontologist, Rawls bases his notion of a just society on its members' *duty* to participate in its construction and maintenance through social cooperation: "The most important natural duty is that to support and to further just institutions" [14, p. 293]. It is thus fundamental to Rawls' theory that it is the *members* of a society that cooperate socially to construct, maintain, and uphold the institutions that ensure a just and *well-ordered* democratic society [14, p. 4].The human capabilities that make this social cooperation possible are two *moral powers,* which I will now describe Rawls' conception of. The second of these is the easiest to comprehend. It is a "capacity for the conception of the good," i.e., "what is of value in human life," which Rawls relegates to sets of "religious, philosophical, or moral doctrines" [15, p. 19]. Rawls is thus not prescriptive but focuses on the moral ability to perceive good and bad, however, one's values may define these categories. The first (and slightly more complex) moral power is more directly related to the topic discussed here. It is the "capacity for a sense of justice…to understand, to apply, and to act from (and not merely in accordance with) the principles of political justice that specify the fair terms of social cooperation" [15, pp. 18–19]. Thus, Rawls' entire concept of a just society rests on its members engaging actively in social cooperation, using their moral powers to produce and maintain a *basic structure* defined in a social contract: "The fair terms of social cooperation are to be given by an agreement entered into by those engaged in it." [15, p. 15] For those who may disagree with some of the social contract's stipulations, social cooperation means working to change the contract from inside the *basic structure* rather than tearing it down.[5] Only this way can the *Difference Principle* and other Rawlsian ethics principles be upheld, and societal institutions of governance be justified. In short, Rawlsian theory prescribes that social contract policies governing a society of sentient beings must emerge from deliberation between *those* sentient beings. It simply cannot be dictated by a third party, not even if that third party is an AI agent constructed by the aforementioned sentient beings with the purpose of governing them. An AI agent-produced "algorithmic social contract" that governs humans, as suggested by Ashrafian [10, p. 7] is thus inconsistent with the fundamental conditions of Rawls' theory of justice, simply because those humans are not directly involved in the deliberations that produce the social contract.

It is worth emphasizing that Ashrafian is clear that he wishes the contract to be constructed by an AI agent rather than, for example, by humans who might be assisted by an AI system. He writes that "this interpretation of Rawls offers a solution to be made synthetically by an artificially intelligent agent," drawing a distinction between the "objective capabilities of algorithms" in "the application of…AI" and human-created social contracts that have historically "been subject to multiple irrationalities and biases inherent to human nature" [10, p. 7]. By using language such as "any population *under* an algorithmic AI social contract" [10, p. 5] (italics mine), Ashrafian makes it clear that he proposes a contractarian system in which a society's population *acquiesces* to an "algorithmic AI social contract" produced by an AI agent, which he claims "may offer a more measured selection of choices in a social contract for society and governmental policy" [10, p. 3, 7]. Because Ashrafian's ultimate objective is to let an AI agent produce the social contract, it does not make his proposition more consistent with Rawlsian theory when he grants "any population under algorithmic AI social contract…the ability to choose and change the nature of any underlying algorithm and its applications." Moreover, how would such a change in the "nature of the underlying algorithm" be decided? If the underlying algorithm, as Ashrafian suggests, is based on a formalized version of Rawlsian theory, any adjustment done by humans could potentially lead to less adherence to Rawls' principles. The only way to avoid that is to also decide on the adjustments through an *original position/veil of ignorance* process. But why not, then, skip a step and deliberate over the social contract directly? What is the need for an AI agent in that situation? It seems that Ashrafian's proposition is caught in a theoretical cul-de-sac: on one hand, Rawlsian theory dictates that the agent cannot be left alone to construct a social contract without human deliberation because the social contract will govern those humans. On the other hand, if the agent's "underlying" algorithm(s) can be adjusted by humans, it would require an *original position/veil of ignorance* process to keep it Rawlsian—and that makes the AI agent redundant.

Another fundamental Rawlsian principle that is broken by Ashrafian's proposal is the *free and equal* participation of the deliberating parties in the original position. Ashrafian is clear that his proposal makes use of "AI technology of the current era," which he deems "able to offer an algorithmic

[5] Rawls makes it clear that this only applies to *well-ordered* societies and that breaking the contract through civil disobedience is justified when the basic structure does not allow for change through democratic means, for example, in a dictatorship or even a pre-totalitarian, democratic state [14, p. 319].

social contract that would be at a technological ideal" [10, p. 3]. There is no need, then, to engage with imaginaries related to sentient machines and potential Artificial General Intelligence (AGI) agents in the consideration of Ashrafian's proposition.[6] A more fundamental question is whether *any* form of AI agent can be considered *free and equal*. Per Powers and Ganascia [7] and following Russell and Norvig [9], current implementations of AI agents are *not* free, and thus they cannot be the sole deliberators of a social contract in the Rawlsian sense. But the *free and equal* condition also restricts hybrid processes in which humans deliberate with AI agents to reach a *fair agreement* as a social contract simply because AI agents can never be considered equal to their human counterparts. Embodiment, history, biology, forms, levels of intelligence, and many other factors will always put AI agents and humans on unequal footing.[7]

3 *Extrinsic* and *intrinsic* adherence to Rawlsian theory

The above examination of Ashrafian's proposition shows that a Rawlsian social contract and *basic structure* design cannot solely be the work of an AI agent. Yet, there are other ways in which Rawls' theory of justice as fairness may be applied in an AI context comprehensively. This can happen in at least two apparent ways, which I will classify as *intrinsic* and *extrinsic* adherence to Rawlsian theory. An AI system adheres *intrinsically* to Rawlsian ethics when Rawlsian theory is essential to the AI system's design. An example of this could be a system containing a machine-learning AI agent whose algorithms are trained and conditioned to adhere exclusively to Rawls' theory of justice as fairness. This is different from an AI system with *extrinsic* adherence to Rawlsian principles, which follows Rawls' theory of justice as fairness because external forces compel it to do so. This could be an AI system situated in a zone where the legislative regulation of AI technology adheres to Rawls' principles. There is a notable difference between these two states of adherence. An AI system with *intrinsic* adherence to Rawls can be situated in a legislative zone where laws do not necessarily correspond to Rawls' principles, such as in the United States, where specific AI regulations have yet to emerge at the time of writing. Conversely, an AI system with *extrinsic* adherence *must* be governed by legislation or enforced policies that adhere to Rawlsian principles, such as the EU's AI Act appears to do to some extent (see below). An extrinsically adherent AI system can be designed for more general purposes, even if it is exclusively used in ways that adhere to Rawlsian principles because of the legislation in place.

Incidentally, intrinsic adherence to Rawlsian principles does not necessarily mean that the system is built around the *original position*, *the veil of ignorance*, or other principles tied to the *basic structure.* Such a system could integrate other Rawlsian principles, as exemplified by Zhang and Shah, who propose a mathematical formulation of Rawls' *Difference Principle* [16]. Or it could make use of entirely new frameworks inspired by Rawls, such as those suggested by Heidari et al. [2, 3] and Verdiesen et al. [4]. If an AI agent with intrinsic adherence to Rawlsian principles *does* make use of concepts such as the *original position* and *the veil of ignorance* as part of some sort of social contract generation, it must simultaneously adhere extrinsically to Rawls' principles. That is, there must already be a Rawls-consistent, human-deliberated *basic structure* in place governing the AI agent to avoid an incongruency with Rawlsian theory, at least if the *basic structure* also governs humans. In such a case, the AI agent's use of the *original position* and the *veil of ignorance* concepts must be linked directly to the human-deliberated *basic structure* to avoid violating Rawlsian principles. This is one way in which an AI agent or AI system can support the maintenance of the *basic structure*. In the following, I will explore additional examples of how Rawls' theory can be implemented extrinsically and intrinsically.

4 Extrinsic adherence: the *publicity* criterion and the EU's AI Act

Rawls' *basic structure* is underpinned by what he calls *background institutions.* These are the entities that make society *well-ordered* in Rawls' terminology: "Agreements in everyday life are made in determinate situations within the background institutions of the basic structure" [15, p. 15]. Furthermore, it is the "background institutions that secure the basic equal liberties…as well as fair equality of opportunity" [15, p. 43]. The *background institutions* can be actual, human-led institutions that secure justice, equality, and liberties such as, for example, the judicial institutions and regulatory agencies of a democratically elected government. But it can also be "institutions" such as taxation or property that are conceptual institutions first and societally-implemented institutions second [14, p. 234].

Iason Gabriel provides an example of an extrinsic application of Rawls' theory in the AI context that is both consistent with the theory of justice as fairness and applies to the appropriate *basic structure* level [6]. Gabriel's application

[6] Moreover, it is at least debatable whether such an AGI agent might even possess the *moral powers* necessary to engage in the *original position* process, a point for which I am indebted to one of the article's reviewers.

[7] In this particular context, if an AI agent was equal to a human, would they not simply be considered human, thus making the question redundant?

is extrinsic since it calls for a justice framework that can govern various forms of AI technology implementations, thus establishing the primacy of human deliberation for the *basic structure.* Gabriel focuses on *background institutions* in alignment with scholars who have previously identified AI-related social injustices against Black and brown people, LGBTQ+people, and others [17–19]. Because Gabriel wishes to ignite a theoretical discourse on Rawlsian justice in AI, his paper mostly has a survey-like character. One extrinsic application of Rawlsian theory that Gabriel does explore in some depth is Rawls' concept of *publicity* and how it conflicts with one of the most controversial aspects of real-world AI implementations: the "black box" nature of some AI systems [6, 20]. Corporations offering AI-related products often consider elements such as training data sets and algorithmic design to be trade secrets, fearing the loss of a competitive edge [20]. Gabriel argues that this closedness conflicts with Rawls' *publicity* criterion, which calls not just for public participation in *basic structure* institutionalization but also for *public reason*, i.e., a public deliberation that legitimizes the institutions. For this to happen, it is obvious that the public must have access to information about the institutional mechanisms. This means that if AI systems are used in government functions or in situations where private, commercial actors provide public or public-like services,[8] Rawlsian theory requires that the systems must be subject to the highest degree of public transparency possible without trading off national security. Commercial companies are disincentivized by market competition to provide this level of transparency and must likely be compelled to do so by *background institutions* in the *basic structure.*[9]

Gabriel presents a convincing (albeit brief) argument that an extrinsic application of the Rawlsian principle of *publicity* could, for example, compel public auditability of what would otherwise be considered black-boxed trade secrets in the case of AI systems that are in public or public-like use, including any governmental or municipal use of AI. Notably, public auditability is not the only way to achieve the level of transparency inherent in the Rawlsian *publicity* concept. The AI Act moving through the European Union's legislative process at the time of writing has several provisions that support Rawlsian *publicity* without necessitating full, public audits. For example, it demands that so-called "Limited risk AI systems should comply with minimal transparency requirements that would allow users to make informed decisions." Generative AI systems are required to "publish summaries of copyrighted data used for training," and the EU will assess all high-risk AI systems before they enter "the market and also throughout their lifecycle." The latter will be accompanied by registrations in publicly-accessible EU databases, either in a specific high-risk AI system database or already-established databases related to the EU's product safety legislation [22]. Although the United States has yet to enact similar legislation, the Biden-Harris administration's Office of Science and Technology Policy has proposed a "blueprint" for an *AI Bill of Rights* that may eventually turn into legislation, and as a stop-gap measure, the administration has "secured voluntary commitment" to AI safety from seven leading providers of AI systems [23]. The commitments include "internal and external security testing of their AI systems before their release," "sharing information across the industry and with governments, civil society, and academia on managing AI risks," "facilitating third-party discovery and reporting of vulnerabilities in their AI systems," "publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use," and "prioritizing research on the societal risks that AI systems can pose, including on avoiding harmful bias and discrimination, and protecting privacy" [23]. Though not legally binding and too general to be highly effective, these commitments are all examples of extrinsic applications of Rawls' *publicity* principle that do not necessarily require public audits.

Rawls leaves room for some withholding of secrets but acknowledges that it is a matter of a trade-off between *publicity* and his support for a "property-owning democracy" [15, p. 139], which seems to encompass the necessity of the right to trade secrets if technological innovation is to continue. On the other hand, Rawls also insists that a *well-ordered society* is "not…a private society" [15, p. 199]. Rawls' position may thus be interpreted as allowing for private ownership of, for example, trade secrets, as long as these do not have consequences or implications for the *basic structure* or its derived *background institutions*. Consider what Rawls writes about *publicity*: "Publicity ensures, so far as practical measures allow, that citizens are in a position to know and to accept the pervasive influences of the basic structure that shape their conception of themselves, their character and ends." In other words, it matters in which situations the *publicity* criterion is applied. This is reflected in the EU's AI Act and its division of AI systems into four levels of risk attributable to AI systems: Unacceptable, High-Risk, Limited Risk, and Minimal Risk. The AI Act demands transparency corresponding to Rawlsian *publicity* in the cases of High-Risk and Limited Risk AI systems, a rule which both addresses and extends beyond Rawls' *basic*

[8] "Public-like" examples include AT&T's government-sanctioned monopoly on telephony in the United States until circa 1901–1984, Google and Apple's provision of a contact-tracing infrastructure to governments around the world during the COVID-19 pandemic, and Meta's Facebook platform with its 2bn + user base.

[9] It should be acknowledged here that the open-source AI movement puts pressure on commercial companies to choose a strategy of openness despite market interests. The movement has achieved some, though not universal, success in this matter [21]

structure and its *background institutions*. The AI Act also permits "regulatory sandboxes" in which the "development, training, testing, and validation of innovative AI systems" can be performed under government supervision [8].

5 Intrinsic adherence: the epistemology of prediction and Rawls' critique of utilitarianism

Under intrinsic adherence to Rawls' theory, his principles are set as boundaries or parameters for, e.g., reasoning in an AI system used in advisory functions. Since, as argued above, an AI agent cannot singlehandedly enter into deliberations in the *original position* without violating Rawlsian principles, the following examples of intrinsic adherence describe AI systems of an assistive kind, and if any AI agents are involved, they are not in any positions where they can singlehandedly govern humans. Widespread use cases of this kind are AI systems that provide decision-making recommendations based on modeling and prediction or AI agents that semi-automate decision-making based on similar prediction mechanisms. However, as I will show in the following, there are some strong incongruencies between Rawlsian ethics principles and how some AI systems enable decision-making through prediction. Since most prediction functions in AI systems are probabilistic, basic statistical boundary conditions are in place, including what and how much is measured and the temporal reach of inferences. The kind of ethical *basic structure* policymaking that Rawls dedicated his life to theorizing rarely concerns the short term, and thus, AI agents that operate on a short time scale (e.g., in autonomous vehicle operation or automated high-frequency trading systems) are considered outside the scope of this article. Beyond the temporal dimension, *basic structure* considerations concern large groups rather than individuals by definition. Hence, the following discussion will be focused on the former.

A school of ethics in which predictions about benefits for large groups are particularly relevant is *utilitarianism.* Rawls is known for his opposition to utilitarianism, which he argues can be used to justify abhorrent conditions in society such as slavery and extreme oppression of minorities. Furthermore, Rawls argues that utilitarianism is *teleological* precisely because of its predictive nature. Utilitarianism's core objective, maximizing the good for as much of society as possible, involves predicting the consequences of decisions and actions, which is why utilitarianism is characterized as a *consequentialist* theory. For this reason, utilitarians like Rawls' critic Harsanyi rely on probabilistic theory to justify their ethics [24], reflecting the post-WWII rise of epistemic probabilism owing to the increase in available computing power. Later came the realization that humans are irrational at times, making it much more difficult to predict anything involving human behavior than, say, predicting biological occurrences. This was later followed by a well-founded skepticism toward the epistemological value of analyzing large datasets [25, 26]. Does that mean that we should never consider the consequences at all when making large-scale decisions? This approach is not viable either according to Rawls, who writes that "all ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy…" [14, p. 26]. Therefore, while Rawls does not agree with the utilitarian standpoint that our ability to make probabilistic estimations of the consequences of an action or decision sufficiently enables moral judgments, he also does not want to completely forego consideration of potential consequences in ethical decision-making. This raises the question of the *extent* to which Rawls accepts consideration of anticipated consequences. Which predictions and anticipations are valid for Rawls, and which are not?

This question is best answered illustratively by visiting the work on prediction done by Karl Popper [27]. Using his characteristic, good-natured polemicism, Popper distinguishes between *scientific prediction* and *unconditional historical prophecies* [27, p. 456], ascribing validity to the former and calling the latter "superstition" [27, p. 459]. The difference, argues Popper, is that scientific prediction is concerned with matters that are *conditional*, i.e., related to consequences of changes to phenomena that tend to be stable in their behavior over the long term. For example, it would take a significant natural event for the Earth to stop rotating and for the celestial bodies in our solar system to stop their heliocentric movements. For that reason, the confidence with which we can predict that the sun will rise tomorrow is so high that it is near-certainty. Because of the relative regularity of the Earth's rotation, we can also predict *when* the sun will rise. This means that winemakers are able to predict with greater confidence when their grapes will be ready to harvest. But human behavior is not as conditional as that of grapes. Human behavior, rather, is determined by a multitude of variables, and we do not often remain in systems that are (in Popper's description of the required conditions for human predictability) "isolated, stationary, and recurrent." "These systems are rare in nature," Popper continues, "and modern society is surely not one of them" [27, p. 457]. Popper does not completely reject predictions about humans, of course. He merely makes a strong case for the implausibility of high-confidence, long-term predictions about humans based on irregular historical events and theories that are not grounded in convincing amounts of empirical data with strong conditionality. He even acknowledges that certain human behavioral patterns are somewhat conditional: "We can learn from the economist that under certain social conditions, such as shortage of commodities, controlled prices,

and, say, the absence of an effective punitive system, a black market will develop" [27, p. 456].

Nevertheless, Popper argues that two characteristics severely diminish the value of predictions about human behavior: irrationality and knowledge acquisition. The human tendency to abandon rational choice at seemingly unpredictable times has long been an Achilles' heel for predictive systems [28, 29] including game theory (as used in Ashrafian's proposition), which has also been criticized for not accounting sufficiently for interactivity [30]. Yet, it is Popper's second contention that remains the most powerful in the present context. He contends that we cannot know with *any* degree of certainty what kind of knowledge we will develop through our scientific endeavors and how this may impact human behavior. History is full of unexpected discoveries, but even more full of unintended consequences of applications of newly generated knowledge, he argues. For Popper, this is a situation completely unguided by conditionality and, therefore, unpredictable. In his view, it is the role of the social sciences to try to anticipate any unintended consequences of human actions and decisions through theory. While theory is not fact, it does not claim to be prediction either, and Popper sees it as useful, well-argued conjecture and speculation based on well-founded, previously established knowledge [27, pp. 454–455].

6 AI prediction versus AI prophecy

This brings us back to Rawls' distinction between deontology and utilitarianism. As mentioned, Rawls sees utilitarianism as teleological, while "deontological theories are defined as non-teleological ones" [14, p. 26]. For Rawls, then, the ethics of just and fair decision-making at the societal level cannot be tied to predictions of future consequences of that decision because, as Popper showed above, we cannot *know* what the consequences of human actions are beyond conditional situations happening very close to the present. We can predict the latter with sufficient probability in the short term, but any statement of consequences occurring further into the future are merely, as Popper calls them, *prophecies.* Rawls agrees with Popper about the untrustworthy nature of predictions of human behavior, writing that "the general capacities of mankind" are unknown [14, p. 184]. Furthermore, Rawls describes an epistemological differentiation similar to Popper's by stating that parties working to create a just *basic structure* in the original position face *uncertainty* rather than *risk* as they make their decisions. Risk, he writes, has "some objective evidential basis for estimating probabilities, for example, relative frequencies, or actuarial tables, or the relative strengths of the various propensities of things (states of affairs) that affect the outcome." With uncertainty, on the other hand, "there is no such objective basis; such bases as there may be are highly intuitive and sketchy" [15, p. 106]. Thus, Rawls' concept of uncertainty corresponds roughly to Popper's *prophecies* while the "evidential basis" involved in risk assessment can be viewed as a cousin of Popper's conditional facts.

Rawls makes good use of the epistemological weakness of uncertainty in the *original position*. By design, the parties working to construct a just *basic structure* "have no reliable basis for estimating the probabilities of the possible social and historical conditions, or the probability that the persons they represent affirm one comprehensive doctrine (with its conception of the good) and not another" [15, p. 106]. It is essential, then, that probabilities cannot be estimated in the *original position,* as their absence ensures that the deliberating parties exclusively discuss justice for the people involved in the situation at hand. The moral fabric of a society's *basic structure* can thus not be determined by assumptions about how society *might* develop. Economists and climate scientists may be able to make reasonably confident projections of how some phenomena will develop in the near-term future, but per Popper, the presence of the human factor can turn these projections into mere *superstition.* At the large-group scale, for example, few economists predicted that people would turn so massively towards working remotely through the COVID-19 pandemic that overall household spending stayed up (while economic inequality increased) and resulting in post-pandemic inflation and the so-called *great resignation* [31–33]. Similarly, it remains very difficult to predict the reactions of large groups of humans to truly destructive climate change effects, such as migration patterns, with any significant confidence [34]. These are just two examples of events governed by the type of societal structure Rawls' theory is concerned with at the large group level.

Thus, if Rawls and Popper are right, Rawlsian ethics would find that predictions about events contingent on human behavior are bound to be too speculative for making decisions about large-scale societal issues that impact individuals. And yet, AIs are routinely employed to make precisely these kinds of predictions in the *background institutions* of the *basic structure.* One example is when AI systems perform load forecasting in a nation's or a city's energy infrastructure or other critical infrastructure, such as water delivery. These are cases in which the implications of poor prediction accuracy can range from the perpetuation of socioeconomic injustice to fatalities during extreme weather or natural disasters [35, 36]. The same principles should apply to any *basic structure* resource allocation where AI systems make recommendations, whether the context is health care, national security, etc.

Overall, Rawls' and Popper's thoughts about using human behavior as an epistemological foundation for any decision-making are worth considering in relation to most AI contexts. In machine-learning systems, for the algorithms

involved to be effective post-training, it must be assumed that the behaviors registered in the training data are bound to be repeated by other humans. It is well-established that training data that are unrepresentative of the population about which the algorithms make determinations and predictions will cause the latter to be untrustworthy and useless [37]. But this is also the case if it cannot even be expected that humans follow the same patterns of behavior as those in a given training data set, which potentially diminishes the quality of *any* prediction about human behavior produced by machine-learning systems. This quandary resembles Hume's 1748 *problem of induction* [38], which Popper also discusses.

7 The *good* and the *right:* utilitarianism in AI ethics

On the other hand, it could be argued that this simply means that an AI-generated prediction will not *always* be correct. It could still be correct most of the time, thus justifying its use. This is a utilitarianist[10] position often taken in the technology industry, particularly in Silicon Valley. The teleology of assuming an AI is correct in its determinations most of the time contrasts with the overall deontological position held by Rawls. As Rawls describes it, in teleological approaches, "the good is defined independently from the right" [14, p. 22], and in utilitarianism, it is viewed as *right* to maximize the *good*. One of the core problems of this, according to Rawls, is that when the *good* is left morally undefined or relative, "it enables one to judge the goodness of things without referring to what is right" [14, p. 22]. In other words, in utilitarianism, it can be seen as *right* to justify the maximization of something if that something is deemed *good* by either a governing body or a majority of the people. This is how utilitarianism can be used to justify the slavery practiced in the early history of the United States and its pre-revolution colonies, per Rawls [14, p. 145].

The AI case is similar in that the *good* is determined by a small group of people who justify its use by maximizing it (because such action is viewed as *right*) in accordance with utilitarianism. There is a Rawlsian point to be made about how a utilitarian approach allows for such a small group of people to codify data as *good* on behalf of large populations, but this lies outside the scope of the present paper and deserves a much more extended analysis. Instead, let me turn to a real-world example, namely the surge of popularity seen by generative AI products in the 2022–2023 timeframe. Microsoft was quick to integrate OpenAI's GPT 3.5/4 models into widely used products, including the search engine Bing. It was likely assumed by Microsoft that this would be a useful addition to their products which would raise productivity among the users (the *good*), and so the company rushed to maximize the availability of it (the *right*), maintaining a competitive advantage at the same time. But even though Microsoft holds massive shares of its markets, it can hardly be called a *background institution*, and as such, Rawls' *basic structure* principles do not necessarily apply.[11] The principles *do* apply, however, to Google, which has a near-monopoly on search engine-based information retrieval, as well as an overwhelming share of the smartphone OS market through its Android platform. As van Dijck et al. argue, this gives the company a status comparable to (or lets it provide the platform and infrastructure for) public services and utilities [41], making it an example of a privately held *background institution.* In 2023, Google rushed to follow Microsoft's head start on the integration of generative AI in consumer products, causing concern among the employees that the product testing would be insufficiently performed prior to launch [42].

In their risk assessments, both Google and Microsoft appear to have assumed, teleologically, that the harms related to such a rollout could be contained, thus circumventing a discussion of their conception of the *good*, i.e., the overall benefits of launching the products. By maximizing what it sees as benefits for both the company and its customers, it does what is *right*, according to utilitarianism. This is exactly what Rawls uses as an argument *against* utilitarianism. At the time of writing, what harms may emerge from the expedited rollouts of generative AI technologies in consumer products from organizations holding near-monopolies, such as Google, is unknown. Only when any harms occur can it be known whether the companies made the right choices. Hence, their decisions were teleological. More importantly, if harm comes to a smaller group of people than those who benefit, the companies would still be able to justify their decisions through utilitarianism.

Deontologists such as Rawls instead focus on the *good* itself—and ask, is it fair and just? In this case, from a Rawlsian perspective, the question to ask is whether it is *fair* that some people are harmed by a particular AI implementation, even if it can be argued that a majority benefit from it. Buolamwini and Gebru famously established that facial recognition algorithms in a range of AI systems struggled with the recognition of people of color [43]. A utilitarian view would argue that as long as the system works for a majority of people, it is justifiable to roll it out and let people of color wait until the system can be adjusted along the way.

[10] It is not merely a utilitarian position but can be more or less seen as an ideologically-determined view, a belief in the superiority of utilitarianism [39, 40]

[11] I am indebted to an anonymous reviewer for this distinction.

A deontological (and thus Rawlsian) view would hold that such inequitable treatment is unfair, unethical and should not occur because it reinforces inequities and power asymmetries already in place. A Rawls-inspired view might be, then, that any AI that is rolled out widely must be thoroughly tested for the worst possible outcomes before it is released, even if this means risking a competitive edge or slower development of the technology. This is, incidentally, one of the requirements of the EU's AI Act, at least when it comes to "high-risk" AI systems such as those implemented in law enforcement, education, critical infrastructure management, access to public services, etc. It is also a core tenet of the Biden-Harris administration's AI regulation blueprint.

8 Conclusion

Above, I have elucidated the complexity of comprehensively applying Rawls' theory of justice as fairness to AI systems and AI agents. Using Ashrafian's proposition as an example, I have shown how applying Rawls' theories without engaging fully with his scholarship can lead to theoretical cul-de-sacs and inconsistencies. This is not to say that Rawlsian theory, including central concepts such as the *original position* and the *veil of ignorance,* cannot be applied to AI ethics at all. Rather, the argument stated here is that scholars wishing to do so must either take Rawlsian concepts out of their original context and build something new from them or apply Rawls' theory comprehensively and in accordance with its internal logic and conditions. This means, for example, that Rawls' theory of justice as fairness cannot be used to argue for (or in the production of) a social contract generated by an AI agent, since Rawls argues that those who are governed by such a contract have a moral duty to participate in the creation and maintenance of the contract and the *basic structure* that emerges from it. One contribution of this article to the AI ethics discourse, then, is establishing that AI agent-generated social contracts cannot be defended through Rawls' theory.

Another contribution is the conceptualization of *intrinsic* and *extrinsic* adherence to Rawlsian principles in AI systems, i.e., the difference between when Rawlsian theory governs an AI system as part of its design or as part of the legislative framework within which it is situated. Given the popularity of Rawls' concepts of fairness in AI ethics [1], such a distinction should be useful in the current discourse, as it adds some clarity to the difference between the effects of imposing AI ethics at the legislative level and at the design level. Finally, I have provided a discussion of how Rawls' opposition to utilitarianism plays into AI ethics, at least when it comes to *basic structure*-relevant implementations of AI technology. By combining Rawls' counterarguments to utilitarianism and Popper's theory of prediction into an epistemological argument, I have shown that it violates Rawlsian ethics to employ any large-scale, long-term predictions made by AI systems about human behavior in decision-making related to the *background institutions* of the *basic structure*. Overall, I conclude that policymakers and regulators who wish to adhere to Rawls' theory of justice as fairness must be highly selective when choosing the contexts in which they apply AI systems and agents in their work. More than anything, however, the above article is a furtherance of the discussion of the role played in AI ethics by John Rawls' work, and thus also an encouragement for further development of Rawlsian AI ethics theory by others.

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Declarations

Conflict of interest No conflicts of interest.

References

1. Jørgensen, A.K., Søgaard, A.: Rawlsian AI fairness loopholes. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00226-9
2. Heidari, H., Ferrari, C., Gummadi, K., Krause, A.: Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 1265–1276. Curran Associates, Inc., New York (2018)
3. Heidari, H., Loi, M., Gummadi, K. P., Krause, A.: A moral framework for understanding fair ml through economic models of equality of opportunity. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 181–190 (2019)
4. Verdiesen, I., Dignum, V., Hoven, J.V.D.: Measuring moral acceptability in E-deliberation: A practical application of ethics by participation. ACM Trans. Internet Technol. TOIT **18**(4), 1–20 (2018)
5. Santoni de Sio, F., Almeida, T., van den Hoven, J.: The future of work: freedom, justice and capital in the age of artificial intelligence. Crit. Rev. Int. Soc. Polit. Philos. (2021). https://doi.org/10.1080/13698230.2021.2008204
6. Gabriel, I.: Toward a theory of justice for artificial intelligence. Daedalus **151**(2), 218–231 (2022)
7. Powers, T.M., Ganascia, J.-G.: The ethics of the ethics of AI. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The Oxford Handbook

of Ethics of AI, pp. 25–51. Oxford, Oxford University Press (2020)
8. Madiega, T. A.: Artificial intelligence act. European Parliament, Briefing PE 698.792, Jun. 2023.
9. Russell, S.J., Norvig, P.: Artificial Intelligence - A Modern Approach, 4th edn. Pearson Education, Inc., Boston (2020)
10. Ashrafian, H.: Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00253-6
11. Chen, V., Hooker, J. N.: A just approach balancing Rawlsian leximax fairness and utilitarianism. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 221–227 (2020)
12. Grace, J., Bamford, R.: 'AI theory of justice': Using Rawlsian approaches to legislate better on machine learning in government. Amic. Curiae **1**, 338 (2019)
13. Rawls, J.: Collected Papers. Harvard University Press, Cambridge (2001)
14. Rawls, J.: A Theory of Justice – Revised Edition. Belknap Press, Cambridge (1999)
15. Rawls, J.: Justice as Fairness: A Restatement. Harvard University Press, Cambridge (2001)
16. Zhang, C., Shah, J.: On fairness in decision-making under uncertainty: Definitions, computation, and comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence (2015)
17. Le Bui, M., Noble, S.U.: We're missing a moral framework of justice in artificial intelligence. In: Dubber, M.D., et al. (eds.) Oxford Handbook of Ethics AI, pp. 163–179. Oxford University Press, Oxford (2020)
18. Benjamin, R.: Race after Technology: Abolitionist Tools for the New Jim Code. Oxford Univ. Press, Oxford (2019)
19. Noble, S.U.: Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, New York (2018)
20. Brevini, B., Pasquale, F.: Revisiting the Black Box Society by rethinking the political economy of big data. Big Data Soc. **7**(2), 1–4 (2020)
21. Ackermann, R.: The future of open source is still very much in flux. *MIT Technology Review*, Aug. 17, 2023. https://www.technologyreview.com/2023/08/17/1077498/future-open-source/ (2023). Accessed 19 Aug 2023
22. European Parliament: EU AI Act: first regulation on artificial intelligence | News | European Parliament, Aug. 06, 2023. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (2023). Accessed 18 Aug 2023
23. The White House: FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, *The White House*, Jul. 21, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ (2023). Accessed 19 Aug 2023
24. Harsanyi, J.C.: Bayesian decision theory, rule utilitarianism, and Arrow's impossibility theorem. Theory Decis. **11**(3), 289–317 (1979). https://doi.org/10.1007/BF00126382
25. Tversky, A., Kahneman, D.: Rational choice and the framing of decisions. J. Bus. (1986). https://doi.org/10.1086/296365
26. Hong, S.: Technologies of Speculation: The Limits of Knowledge in a Data-Driven Society. New York University Press, New York (2020)
27. Popper, K.: Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, New York (2014)
28. Boudon, R.: Beyond rational choice theory. Annu. Rev. Sociol. **29**(1), 1–21 (2003)
29. Hodgson, G.M.: On the limits of rational choice theory. Econ. Thought **1**(1), 2012 (2012)
30. Colman, A.M.: Cooperation, psychological game theory, and limitations of rationality in social interaction. Behav. Brain Sci. **26**(2), 139–153 (2003)
31. "WDR 2022 Chapter 1. Introduction," World Bank. https://www.worldbank.org/en/publication/wdr2022/brief/chapter-1-introduction-the-economic-impacts-of-the-covid-19-crisis. Accessed 26 May 2023
32. Cascaldi-Garcia, D., Orak, M., Saijid, Z.: Drivers of Post-pandemic Inflation in Selected Advanced Economies and Implications for the Outlook, Jan. 2023. https://www.federalreserve.gov/econres/notes/feds-notes/drivers-of-post-pandemic-inflation-in-selected-advanced-economies-and-implications-for-the-outlook-20230113.html, Accessed 26 May 2023
33. Fuller, J., Kerr, W.: The Great Resignation Didn't Start with the Pandemic. *Harvard Business Review*, Mar. 23, 2022. https://hbr.org/2022/03/the-great-resignation-didnt-start-with-the-pandemic, Accessed 26 May 2023.
34. Bryne, C.: Climate change and human migration. UC Irvine Rev **8**, 761 (2018)
35. McGovern, A., Ebert-Uphoff, I., Gagne, D.J., Bostrom, A.: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. Environ. Data Sci. **1**, e6 (2022)
36. Walker, G.: Environmental justice, impact assessment and the politics of knowledge: the implications of assessing the social distribution of environmental outcomes. Environ. Impact Assess. Rev. **30**(5), 312–318 (2010)
37. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, 1st edn. Crown, New York (2016)
38. Hume, D.: An Enquiry Concerning Human Understanding and Other Writings. Cambridge University Press, New York (2007)
39. Srinivasan, A.: Stop the robot apocalypse: the new utilitarians. Lond. Rev. Books **37**(18), 3–6 (2015)
40. Healey, K., Woods, R.H.: Processing is not judgment, storage is not memory: a critique of Silicon Valley's moral catechism. J. Media Ethics **32**(1), 2–15 (2017). https://doi.org/10.1080/23736992.2016.1258990
41. Van Dijck, J., Nieborg, D., Poell, T.: Reframing platform power. Internet Policy Rev. **8**(2), 1–18 (2019)
42. Grant, N., Weise, K.: In A.I. Race, Microsoft and Google Choose Speed Over Caution. *The New York Times*, Apr. 07, 2023. https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html, Accessed 20 Aug 2023.
43. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, PMLR, pp. 77–91 (2018)