Rawls's Original Position and Algorithmic Fairness

Ulrik Franke^{1,2}

Received: 9 August 2021 / Accepted: 19 October 2021 / Published online: 3 November 2021
© The Author(s) 2021

Abstract

Modern society makes extensive use of automated algorithmic decisions, fueled by advances in artificial intelligence. However, since these systems are not perfect, questions about fairness are increasingly investigated in the literature. In particular, many authors take a Rawlsian approach to algorithmic fairness. This article aims to identify some complications with this approach: Under which circumstances can Rawls's original position reasonably be applied to algorithmic fairness decisions? First, it is argued that there are important differences between Rawls's original position and a parallel algorithmic fairness original position with respect to risk attitudes. Second, it is argued that the application of Rawls's original position to algorithmic fairness faces a boundary problem in defining relevant stakeholders. Third, it is observed that the definition of the least advantaged, necessary for applying the difference principle, requires some attention in the context of algorithmic fairness. Finally, it is argued that appropriate deliberation in algorithmic fairness contexts often require more knowledge about probabilities than the Rawlsian original position allows. Provided that these complications are duly considered, the thought-experiment of the Rawlsian original position can be useful in algorithmic fairness decisions.

Keywords Algorithmic fairness · Original position · Veil of ignorance · Veil of uncertainty

1 Introduction

In modern society, automated algorithmic decisions are everywhere. While this has been true for a long time, recent advances in artificial intelligence (AI) have accelerated an existing trend of automation. As an increasing number of tasks—such

✉ Ulrik Franke
ulrik.franke@ri.se

¹ RISE Research Institutes of Sweden, SE-164 29, Kista, Sweden

² KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

as reading and writing texts (see, e.g., Hirschberg & Manning 2015) or identifying objects in images (see, e.g., Zhao et al. 2019)—can be performed by machines, society is gradually transformed.

However, these technical systems are not perfect. In particular, it has been discovered that automated decisions sometimes disadvantage, e.g., poorer people and those from minorities (Nature, 2016), raising questions about fairness. As a result, there is a growing technical and philosophical literature addressing matters of algorithmic fairness. This is not the place to survey the literature in detail, but useful reviews—from different angles—include Lepri et al. (2018), Chouldechova and Roth (2020), and Mitchell et al. (2021).

One strand of this literature concerns what could be termed a Rawlsian approach to algorithmic fairness—the application of Rawls’s seminal 1971 work *A Theory of Justice* to these questions.¹ Indeed, Rawls has been called “AI’s favorite philosopher” Procaccia (2019).

For example, the work by Dwork et al. (2012) and Joseph et al. (2016)² has been described “as a mathematical formalization of the Rawlsian principle of ‘fair equality of opportunity’” Lepri et al. (2018). The same concept is discussed by Lee et al. (2021), who identify it as the philosophical origin of many technically oriented fairness metrics, and also briefly discuss the difference principle. A more detailed treatment and formalization of Rawlsian equality of opportunity in the context of machine learning is given by Heidari et al. (2019). Almost the same authors also propose a technical mechanism to include fairness criteria, inspired by Rawls, as constraints in the optimization problems solved in machine learning training, thus guaranteeing certain corresponding properties of the resulting trained models (Heidari et al., 2018). In the context of autonomous vehicles, Leben (2017) develops an algorithm for trolley-style situations based on the Rawlsian original position and the maximin rule (for a critical discussion of this proposal, see Keeling 2017).

The rest of this article aims to scrutinize the application the original position (Rawls, 1999b, pp. 102–168) to algorithmic fairness and identify some complications arising when attempting such application. The complications thus found should not necessarily be seen as an exhaustive set, but hopefully as a relevant one.

2 Attitudes to Risk

An important feature of Rawls’s theory is that the parties to the original position would apply the risk-averse maximin rule to their choice; aiming to make the least advantaged in society as well off as possible. This is embodied in the difference

¹The 1971 text was considerably revised in 1975 for the first foreign translation, and this revised edition was published in English in 1999. It is this revised edition (Rawls, 1999b) that is used throughout this article.

²The first preprint version of the paper *Fair algorithms for infinite and contextual bandits* by Joseph et al. was titled *Rawlsian Fairness for Machine Learning*, though later versions do not mention Rawls explicitly.

principle, according to which “[s]ocial and economic inequalities are to be arranged [...] to the greatest benefit of the least advantaged” (Rawls, 1999b, p. 266).³

This application of the maximin rule has been contested, perhaps most famously by Harsanyi (1975) who argues that rational individuals in the original position would be expected-utility maximizers, and thus that Rawls's risk-averse difference principle would *not* be chosen.⁴ However, without accepting Harsanyi's objection, Rawls himself explicitly states that “the maximin rule is not, in general, a suitable guide for choices under uncertainty” (Rawls, 1999b, p. 133). On the contrary, the argument for its use in the original position depends on the particular features of that choice—simply put that it is an irrevocable choice with very high stakes.

Somewhat more precisely, Rawls (1999b, pp. 134–135) argues that there are three features of the original position that favor the maximin rule: (i) *ignorance* (as in “decisions under ignorance” as opposed to “decisions under risk”) of (good estimates of) relevant probabilities, (ii) *small upside* in the sense that there is not much need for what can be gained above the minimum that can be secured by the maximin rule, and (iii) *great downside* in the sense that there are very bad consequences of some outcomes. At least in Rawls's later view, the two latter are the more important ones (Freeman, 2019), to be focused on here, but the ignorance feature will be revisited in Section 5.

Thus, before applying the maximin rule or similarly risk-averse principles to matters of algorithmic fairness, it seems reasonable to investigate to what extent decisions about algorithmic fairness share the three features of Rawls's original position. Of course, the original position is “purely hypothetical” (Rawls, 1999b, p. 19), but some real-world positions may share some of its gravity. For example, there is an ongoing debate about whether AI may pose an existential threat to humanity (for some perspectives, see Bostrom 2014; Bundy 2017; Galanos 2019). Thus, high-level AI governance decisions such as the adoption of international conventions, national legislation, or principles of AI development can, with some credence, be said to share the ignorance and great downside features of Rawls's original position, though the potential benefits of AI make it very questionable whether the small upside feature also holds.⁵ Thus, for such decisions, the Rawlsian argument for the maximin rule at least seems applicable, in particular to the extent that these decisions are not revocable. Of course, such momentous decisions are rare compared to the vast majority of decisions about algorithmic fairness—the more detailed ones made on an everyday basis by software engineers. Still, even if few people have high-level AI governance

³For completeness, it should be remarked that the difference principle, important as it is in Rawls's theory, is subordinate to the first principle of justice: “Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all” (Rawls, 1999b, p. 266). Thus, for example, if a distribution of resources in accordance with the difference principle would somehow lead to citizens not being able to see themselves as free and equal anymore, this distribution is precluded.

⁴For a short and pedagogical introduction to the debate between Rawls and Harsanyi, see Resnik (1987, pp. 40–43).

⁵For a concrete example of the difficulty to assess the stakes in this discussion with any semblance of certainty, see Müller and Bostrom (2014) and take note of the differences in opinion among experts.

decisions in mind when discussing algorithmic fairness, such decisions do exist, and should not be ignored. For example, Fjeld et al. (2020) identify 36 documents aiming to identify and codify ethical principles for AI, with an accelerating trend (cf. the timeline compiled by Fjeld et al. on pp. 18–19). The deliberative efforts behind each of these documents resemble Rawls's original position in their aim to delineate reasonable principles to govern an uncertain future, though the parties are not so much ignorant about their own positions as they are about future technological and social developments.

However, as remarked above, most algorithmic fairness decisions are not like that. They are much smaller in scope, in that they do not govern the use of AI in general, but rather just govern particular decision-support systems used in particular products used by particular customers on particular markets. Many of these design decisions are certainly not irrevocable, but take place repeatedly, subject to ferocious competition. Thus, Rawlsian algorithmic fairness should carefully try to distinguish between different kinds of decisions.

2.1 Decisions Where Risk-Aversion Is Appropriate

Above, we identified high-level AI governance decisions as appropriate candidates for Rawlsian algorithmic fairness by virtue of their similarity with Rawls's original position. However, there are also other decisions where risk-aversion seems appropriate, even if we do not need to make the stronger claim that the appropriate form of risk-aversion is exactly the maximin rule.

Risk-aversion is plausible in algorithmic fairness decisions where the stakes involve Rawls's third feature: very bad consequences of some of the outcomes of alternative choices. Indeed, AI is increasingly used in high-stakes decision-making in military applications (Johnson, 2019), healthcare (Obermeyer et al., 2019), court decisions (Dressel & Farid, 2018), and autonomous vehicles (Leben, 2017). Given the fallibility of AI systems, some form of risk-aversion seems prudent in such cases.

However, even if the stakes are lower, risk-aversion might be warranted. Some contexts are by definition risk-averse: anyone buying an insurance (and not intending to make fraudulent claims) is risk-averse, even if the indemnity limit is very small. Thus, for example, in a system designed to classify insurance claims as valid or invalid, it can by definition be safely assumed that all of the customers are risk-averse. Furthermore, if the insurance company is a mutual one, this by definition extends to the owners as well. However, though we may be pleased to have devised this clear-cut example, it must be cautioned that this is the exception rather than the rule (most insurers are commercial rather than mutual, and most cases are not insurance cases anyway).

2.2 Decisions Where Risk-Seeking Is Appropriate

Since algorithmic fairness and AI governance are often discussed in contexts such as those sketched above, it may appear that risk-aversion is most often appropriate. However, when decisions are *not* unique and irrevocable, as in Rawls's original position, but instead recurring or revocable, risk-aversion is often not appropriate.

Are there even cases where risk-seeking may be appropriate? Consider recommender systems—suggesting books to read, movies to watch, music to listen to, museums to visit, etc. A (very) risk-averse system would only suggest what the user is most likely to like. Many of these recommendations are probably ones that the user could have found anyway, though if the system has access to a very broad range of material, it may still be able to recommend some material that is numerically different from what the user already knew of. For example, if you have visited ten museums with impressionist paintings, such a system may know of an eleventh one, and recommend it to you. From the user perspective, such a system may be useful in a limited sense, but this is usefulness *despite* its risk-aversion, not *because* of it: While some recommended material is numerically different from what the user knew of, it will nevertheless seem very familiar. By contrast, a really useful recommender system will take risks—broadening your perspectives even if it means suggesting some things which you turn out to dislike.

The same goes for computer-generated art. Should AI(-assisted) authors, painters, directors, and composers merely aim to please by adopting the risk-averse creative strategy of imitating existing art? Clearly this is fascinating when done for the first time, as the interest in Bach-like chorales (Liang et al., 2017) and Shakespeare-like sonnets (Lau et al., 2020) bears witness to. However, in the longer run such a conservative strategy might make poor use of great potential.

Some scientific endeavors aided by AI are also similar. If AI is used to try to find novel ways to solve problems, it is counterproductive to constrain it to be risk-averse, trying only minute variations of the old approaches. The best approach available to “AI Scientists” may not resemble the scientific process conducted by human scientist” (Kitano, 2021).

A related example is the AlphaGo system which famously defeated human champion Lee Sedol in 2016. It was trained both by learning from human expert moves and by reinforcement learning from playing against itself. The incorporation of human expert moves might be seen as a kind of risk-aversion. However, AlphaGo was later defeated by its cousin AlphaGo Zero, which started training *tabula rasa*, without any human expertise endowed (Silver et al., 2017). Obviously, AlphaGo Zero was a risky bet—it started out as a worse player than AlphaGo, and might never have surpassed it.

The common feature of the artistic and scientific examples is that the great downside feature does not hold: if the book (movie, piece of music, research direction) turned out to be bad, just stop reading (watching, listening to, pursuing) it. Decisions are revocable, in stark contrast to Rawls's original position.

There is also the case of inherently risky activities in which we engage for precisely that reason. Consider an AI system designed to help pick which horses to bet on in horse racing. If such a system is designed with maximin as a fairness constraint, it will invariably tell you not to place any bet at all. Of course, this trivial system will also be completely useless. On the contrary, it seems that to be successful in this domain, the system ought to embrace the risk-seeking context for which it is designed. Perhaps the most appropriate design would be to embed a risk-seeking system within a risk-averse one: first (risk-aversely) setting aside a limited “thrills” budget, then (risk-seekingly) spending that budget. Here, a more complicated architecture than strictly risk-seeking or strictly risk-averse seems appropriate.

2.3 Decisions Where Risk-Neutrality Is Appropriate

There seem to be convincing cases for risk-aversion as well as for risk-seeking in algorithmic decisions. Are there also cases exactly in between, where risk-neutrality is appropriate? The power of risk-neutral maximization of expected value is revealed by its name: as an average over many decisions, it does indeed maximize expected value. The more such decisions there are (so that the average evens out), the more appropriate risk-neutrality becomes.

Of course, the stakes also matter. Consider financial investment. In the long run, higher returns come at the cost of higher risk—therefore, risk-averse investors tend to end up comparably poorer (see, e.g., Bajtelsmit & VanDerhei 1995; Watson & McNaughton 2007, and the literature cited therein). In the short run, however, a risky investment can go either way. Thus, given the high stakes, some risk-aversion is still to be recommended (such as putting aside some money in less risky assets), but the longer the investment horizon, the more risk-neutral we should be. While it leads too far to recommend strict risk-neutrality in financial investment, the maximin rule is far too conservative to be generally applied in algorithmic fairness decisions about decision-support systems for financial investment.

When the stakes are lower and the decisions are repeated, however, strict risk-neutrality appears appropriate. It is not clear that a motion detector turning on the lights as you move into a room, a system enabling a robotic vacuum cleaner to distinguish dust from non-dust, or an image-recognition system designed to replenish empty shelves in a supermarket should be designed with risk-aversion, as opposed to risk-neutrality, as a fairness criterion. Of course, it is perfectly possible to imagine cases where erroneous decisions by these systems can have dire consequences, but given that other systems and reasonable fail-safes work properly, the stakes are indeed quite low. Thus, if any deviation from risk-neutrality is appropriate in these cases, it is rather to embed a risk-neutral system within a moderately risk-averse one (as in the horse racing example).

2.4 Summary of Attitudes to Risk

The maximin rule and the resulting difference principle are key features of Rawls's theory of justice. However, it is not the case that these would emerge from *any* deliberation behind a veil of ignorance. On the contrary, their selection is highly dependent on particular features of Rawls's original position.

We have argued that there are some decisions about algorithmic fairness which resemble Rawls's original position so that the maximin rule might be chosen behind a veil of ignorance—high-level AI governance decisions about international conventions, laws, etc. (Though, of course, other rules might be chosen even then, as Harsanyi reminds us.) However, most decisions about algorithmic fairness do *not* resemble Rawls's original position. When such decisions are deliberated behind a veil of ignorance, we have argued that many kinds of attitudes to risk can be appropriate—risk-aversion, risk-seeking, risk-neutrality, and more complicated combinations.

3 Scope of Stakeholders

The decision to be taken in Rawls's original position is one of utmost gravity. One reason, as discussed in the previous section, is the fact that it is an irrevocable choice with very high stakes. Another reason, however, is its broad scope of stakeholders—all of society. To simplify matters a bit, Rawls has no problem identifying who the stakeholders are, because they are everyone.⁶

Again, as in the previous section, some high-level AI governance decisions such as the adoption of international conventions, national legislation, or principles of AI development may share this broad scope of stakeholders. But clearly, the vast majority of decisions about algorithmic fairness concern a much smaller set of stakeholders. However, this smaller scope should not be confused with ease of delineation. The application of Rawlsian algorithmic fairness hinges on a satisfactory solution to this boundary problem, not present in Rawls's original context.

Consider assessment of the risk of criminal recidivism, one of the most discussed applications in the literature on algorithmic fairness (for instance, it serves as the introductory example in the literature review provided by Chouldechova & Roth 2020 and is even illustrated by a full page artwork on p. 83). In the literature, the problem is typically construed as one of achieving parity of some statistical measure across different groups: fairness on this statistical conception means, e.g., that men and women or members of different ethnic groups have the same probabilities of being classified as probable reoffenders (see, e.g., the tables and figures from Dressel & Farid 2018 or Hamilton 2019). Heidari et al. (2018), in an example illustrating the implementation of their fairness criteria, inspired by Rawls, explicitly define the benefit levels of four stakeholders for the sake of the numerical example: those of false negatives, true positives, true negatives, and false positives. Clearly, this exhausts the set of stakeholders classified as likely reoffenders or not by a classification algorithm. But, equally clearly, it does not exhaust the set of stakeholders in how a society deals with crime in general or offenders' risk of recidivism in particular. Many residents—hopefully most—in a community will never be suspected of crimes and thus will not be subject to the algorithm's classification. They will, however, be subject to its broader effects: as taxpayers they will bear the unnecessary costs of imprisoning those who are unlikely reoffenders, as potential victims of crime they will bear the costs of failing to imprison likely reoffenders, as citizens they will bear the social and political costs of a system that is not perceived to be fair (in different ways). In short, they should also have a say behind the veil of ignorance.

Similar arguments are easily made in other situations: the stakeholders of a recruitment system are not only those recruited (or not), but also their (prospective) colleagues, the customers and owners of the company, etc.; the stakeholders of a system for evaluating insurance claims are not only those whose claims are approved (or not), but also the rest of the insurance collective, the owners of the insurance company (which is identical to the insurance collective if the company is a mutual one),

⁶Somewhat more precisely, they are everyone belonging to one particular people. But there are many peoples in the world, whose freedom and independence should be respected by other peoples. Rawls develops this idea in *The Law of Peoples* (1999a).

the wider community affected by the insurer's risk governance, etc.; the stakeholders of a recommender system for choosing which books to read are not only those receiving the recommendations, but also the authors and publishers of the books recommended (or not), the stakeholders of the technical platform through which the recommendation is delivered, etc.

On the practical side, such considerations are increasingly acknowledged in the practice of responsible AI development. For example, a tool such as the Envisioning Cards,⁷ used within the wider framework of Value Sensitive Design developed by Friedman (1996), includes measures to help design teams identify otherwise easily overlooked stakeholders and consequences (see, e.g., Dexe et al. 2020, for a concrete application of the Envisioning Cards).

However, on the theoretical side, to develop a proper theory of how to apply the concept of the Rawlsian original position to algorithmic fairness, it is clear that this boundary problem needs to be addressed. Who should be included within the set of stakeholders and be, so to say, represented behind the veil of ignorance?

4 Defining the Least Advantaged

According to the Rawlsian difference principle, inequalities are to be arranged so that they are “to the greatest benefit to the least advantaged” (Rawls, 1999b, p. 266). But who are the least advantaged?

Rawls's least advantaged is not a single individual: it is a *group*: “The serious difficulty is how to define the least fortunate group” (Rawls, 1999b, p. 83). Rawls's solution is to select those who are the least fortunate with respect to (i) family and class, (ii) natural endowments, and (iii) fortune and luck, but “all within the normal range” (ibid.), i.e., removing the most extreme cases from consideration. Why is this so? An important part of the answer is that Rawls (1999b, p. 84) does not want to “distract our moral perception by leading us to think of persons distant from us whose fate arouses pity and anxiety.” The difference principle pertains to people who all contribute to society and are entitled to their share because of their contribution, not because of their needs (see also the insightful commentary of Schmitz 2006, p. 189).

Rawls's concern is the basic structure of society. How do these considerations translate into the realm of algorithmic fairness? Consider the case of a system designed to evaluate insurance claims; approving or not approving them. For the sake of the example, assume that it is a mutual company, so that the insured and the owners are the same, and thus constitute all relevant stakeholders.⁸ These can now be partitioned into five groups. First, those who do not make any claims at all, which will in all reasonable cases be the vast majority. Then, among those making claims, true

⁷<https://www.envisioningcards.com>

⁸As noted above, the mutual insurance company example has been chosen for its simplicity, not its representativeness. Most real cases will exhibit even more complicated structures, and identifying the least advantaged may be correspondingly even more difficult.

positives, true negatives, false negatives, and false positives. To make matters as simple as possible, assume actuarially fair premiums, no deductibles, equally sized claims, and the possibility of full compensation. Then, true positives will have suffered loss $-l$ and will be compensated by $+l$; true negatives will have suffered no loss and get no compensation; false negatives will have suffered loss $-l$ but get no compensation; false positives have suffered no loss, but will be compensated by $+l$. In addition, they will all have paid (in advance) a premium $-\pi l$, as will the non-claimants. Note that $-\pi l$ is not constant but is proportional to the number of claims times the ratio $\frac{TP+FP}{TP+FP+TN+FN}$, where the abbreviations denote the number of true or false positives or negatives, respectively. With perfect classification ($FP = FN = 0$), the latter factor is identical to the true prevalence of losses among the claimants. Assuming, more realistically, that perfection cannot be achieved, the least advantaged group becomes the poor false negatives (who end up with $-\pi l - l$). Crucially, they will be worse off under the insurance scheme than in its complete absence (where they would suffer only loss $-l$, without the additional plight of paying an insurance premium which turns out to be useless). Thus, if the false negatives are permitted the standing of the least advantaged group, the insurance scheme (as opposed to no insurance scheme) cannot be motivated by the difference principle. The false negatives would be better off with no insurance scheme at all.

This appears to be the kind of counterproductive conclusions that Rawls, very reasonably, wants to avoid. Better then to consider only three groups of stakeholders: non-claimants, true positives, and true negatives, all of whom end up with $-\pi l$ and are thus better off under the insurance scheme since they prefer this certain premium to the lottery of (0 with probability $1 - \pi$) or ($-l$ with probability π), or they would not buy insurance at all.

The example was chosen for its simplicity. Still, it exhibits a complication with respect to defining the least advantaged. This suggests that to develop a proper theory of how to apply the Rawlsian original position to algorithmic fairness (and indeed more generally), care must be taken to define the least advantaged in an appropriate way.

5 Knowledge About Probabilities

The insurance example in the previous section points to an additional complication. What if the relevant question is not whether to have an insurance scheme or not, but rather pertains to properties of the statistical distribution between the TP, TN, FN, and FP classes within such a scheme? Indeed, such matters are typically considered core questions of algorithmic fairness (see, e.g., Chouldechova & Roth 2020): probabilities such as positive predictive values and false positive or false negative rates are often precisely what decisions are about. This is not to say that these probabilities can be arbitrarily controlled—clearly, they cannot. But, as eminently demonstrated by Heidari et al. (2018) and many others, machine learning problems can be deliberately defined in different ways so as to prioritize different fairness outcomes expressed as conditions on metrics such as positive predictive values and false positive or false negative rates. For example, a decision to define the machine

learning problem in such a way as to prioritize a lower false negative rate will, realistically, entail a higher false positive rate. The converse also holds: defining the problem to prioritize a lower false positive rate will entail a higher false negative rate. But recall that Rawls (1999b, p. 134) assumes ignorance about (good estimates of) relevant probabilities in the original position. While this ignorance is supposed to guarantee impartiality, in algorithmic fairness contexts, it also means that many seemingly very relevant differences between outcomes cannot even be expressed: For example, to continue the insurance example, an option where the number of poor false negatives is a hundred times the number of undeserving false positives cannot be distinguished from an option where the number of false negatives is a hundredth of the number of false positives. And similarly, in other algorithmic decision-making cases such as prediction of criminal recidivism, granting of loan applications, or identification of obstacles in autonomous driving, different statistical distributions over the TP, TN, FN, and FP classes cannot be distinguished, much less deliberated, in the original position.

Given the importance of these statistical properties in algorithmic fairness contexts, however, this poses a serious challenge to the applicability of the original position in these contexts. At least two responses are possible.

First, it may be emphasized that the deliberation in the original position is supposed to decide general principles about the basic structure of society. It is not meant to produce detailed rules for every particular case. According to this response, then, the role of the original position in algorithmic fairness is limited to those high-level AI governance decisions which were described in Section 2. At this level of abstraction, knowledge of particular probabilities is not necessary. Broad ethical principles can be articulated without these.

Second, knowledge of probabilities may be allowed into the original position. This is in essence the key difference between the Rawlsian original position and the alternative proposed by Buchanan and Tullock (1965), which is otherwise strikingly similar to Rawls.⁹ This setup is often called the *veil of uncertainty*, as opposed to the Rawlsian *veil of ignorance*, to highlight the difference, and the similarities, between

⁹“Recall that we try only to analyze the calculus of the utility-maximizing individual who is confronted with the constitutional problem. Essential to the analysis is the presumption that the individual is uncertain as to what his own precise role will be in any one of the whole chain of later collective choices that will actually have to be made. For this reason he is considered not to have a particular and distinguishable interest separate and apart from his fellows. This is not to suggest that he will act contrary to his own interest; but the individual will not find it advantageous to vote for rules that may promote sectional, class, or group interests because, by presupposition, he is unable to predict the role that he will be playing in the actual collective decision-making process at any particular time in the future. He cannot predict with any degree of certainty whether he is more likely to be in a winning or a losing coalition on any specific issue. Therefore, he will assume that occasionally he will be in one group and occasionally in the other. His own self-interest will lead him to choose rules that will maximize the utility of an individual in a series of collective decisions with his own preferences on the separate issues being more or less randomly distributed.” Buchanan & Tullock (1965, p. 78)

the two concepts. For some further reflections about how the two approaches relate to each other, see Buchanan (1972).¹⁰

Thus phrased, we can identify several tenable positions: (i) restrict the use of the original position to AI governance decisions only, and use a veil of ignorance, (ii) use the original position for AI governance decisions and detailed algorithmic fairness decisions alike, but use a veil of ignorance for the former and a veil of uncertainty for the latter, or (iii) use the original position for AI governance decisions and detailed algorithmic fairness decisions alike, and use a veil of uncertainty for both. Arguments can be made in favor of each of these. However, it is clear that if the original position is to be used for detailed algorithmic fairness decisions where statistical properties of algorithms matter, a veil of ignorance, which excludes knowledge of probabilities, cannot enable the required deliberation.

6 Conclusions

Rawls's thought experiment of the original position is an intellectually attractive one.¹¹ Thus, it is no surprise that, in addition to its original application within political philosophy, it has also been applied within many other fields, including algorithmic fairness problems.

However, as pointed out by an anonymous reviewer, the fact that the original position is an intuitive notion for algorithmic fairness is by itself a rather weak reason to believe that it is in the end also an intellectually useful tool. It is instructive to consider Rawls's argument justifying the original position: "There is however, another side to justifying a particular description of the original position. This is to see if the principles which would be chosen match our considered convictions of justice or extend them in an acceptable way" (Rawls, 1999b, p. 17). Rawls goes on to explain the notion of a wide reflective equilibrium where those considered judgments concerning justice which we are most certain about (e.g., "that religious intolerance and racial

¹⁰As pointed out by an anonymous reviewer, admitting knowledge of probabilities is a relaxation of the conditions of the original condition which resembles Rawls's *four-stage sequence*: Rawls (1999b, pp. 171–176) describes four successive stages—original position, constitutional convention, legislature, and application of rules to particular cases—where the restrictions on available knowledge are gradually lifted, so that in the last stage, everyone has access to all facts. Thus, while the veil of ignorance is in effect in the original position, something like a veil of uncertainty may be in effect at later stages. However, at least with respect to Rawls exegesis, too much must not be made of this resemblance: in a footnote Rawls (1999b, p. 173) explicitly points to the importance of distinguishing his own "moral theory" from "social theory" such as that of Buchanan and Tullock (1965).

¹¹Dworkin (1978) describes it as follows: "I shall assume, then, that there is a group of men and women who find, on reading Rawls, that the original position does strike them as a proper 'intuitive notion' from which to think about problems of justice, and who would find it persuasive, if it could be demonstrated that the parties to the original position would in fact contract for the two principles he describes. I suppose, on the basis of experience and the literature, that this group contains a very large number of those who think about justice at all, and I find that I am a member myself." Dworkin (1978, p. 159)

discrimination are unjust,” p. 17) serve as “provisional fixpoints which we presume any conception of justice must fit” (p. 18). To find the appropriate original position, we then move back and forth: find plausible conditions, see if they can generate significant principles and if these principles match our provisional fixpoints, and when faced with discrepancies, sometimes alter the conditions of the original position, sometimes alter our judgments on justice, until “at last our principles and judgments coincide” (p. 18). (For fuller treatments of reflective equilibrium, see Daniels 2020; Rawls 1999b, pp. 42–45.)

But if this is true with regard to Rawls’s theory of justice in society, something similar is probably also true of a Rawlsian theory of algorithmic fairness. The problem is that even though these questions have received much attention in the past few years, our considered judgments about algorithmic fairness are much less certain than those about social justice. In other words, the provisional fixpoints of algorithmic fairness are much more provisional, much less fixed. This observation should make us cautious in applying the original position to algorithmic fairness problems—there are complications (some of which are resolvable challenges, some of which are limits which cannot be crossed) which must be considered before the original position can reasonably be applied to such problems. Identifying some such complications has been the aim of the preceding sections. Indeed, our findings will hopefully contribute to the process of eventually reaching a reflective equilibrium on matters of algorithmic fairness.

More precisely, we have argued in Section 2 that there are important differences between Rawls’s original position and a parallel algorithmic fairness original position with respect to risk attitudes. Rawls describes three features of his original position—ignorance about probabilities, small upside, and great downside—which favor the maximin rule, which in turn serves as an important justification of the difference principle. While some high-level AI governance decisions share the ignorance and great downside features (though probably not the small upside), most algorithmic fairness decisions are not at all like that, sharing at most a single one of the three features. Instead, we have argued for a more nuanced view of attitudes to risk in algorithmic fairness decisions—finding cases where risk-aversion, risk-seeking, or risk-neutrality, respectively, are reasonable.

Further, we have argued in Section 3 that the application of Rawls’s original position to algorithmic fairness faces a boundary problem, not present in Rawls’s own application of the concept, in defining relevant stakeholders. This observation merits further theoretical investigation, aiming to develop a principled position on which stakeholders should be included, and how to find them.

There is also, as we have observed in Section 4, a complication with respect to defining the least advantaged. Applying the concept of the original position to algorithmic fairness decisions, this must be carefully deliberated, since different definitions of the least advantaged may entail considerably different outcomes.

Finally, we have argued in Section 5 that the ignorance about probabilities inherent in Rawls’s veil of ignorance is problematic in algorithmic fairness contexts. More precisely, differences in statistical properties which are typically assumed to be core questions of algorithmic fairness cannot be expressed and used to distinguish between options if knowledge about probabilities is ruled out. Therefore, Buchanan

& Tullock's veil of *uncertainty*, which admits knowledge about probabilities, is more appropriate than Rawls's veil of *ignorance* if the original position is to be used for detailed algorithmic fairness decisions.

To summarize, the Rawlsian original position can be applied to problems of algorithmic fairness, provided that the following complications are duly considered: Orthodox application—entailing the maximin rule and the difference principle—is reasonable only for a comparatively small set of problems of high-level AI governance decisions.¹² For more detailed decisions on algorithmic fairness, variations on the original position can also be employed, but this requires careful analysis of at least (i) attitudes to risk, (ii) the scope of stakeholders, (iii) the definition of the least advantaged (in the case where the deliberation yields something like the difference principle), and (iv) knowledge of probabilities. Provided that these complications are duly considered, the thought-experiment of the Rawlsian original position can be useful in algorithmic fairness decisions. However, it must not be assumed that Rawls offers any simple algorithm which can be automatically applied out of its original context in order to guarantee fairness. Additional deliberation is needed to reach a reflective equilibrium on matters of algorithmic fairness.

Funding Open access funding provided by RISE Research Institutes of Sweden. This work was supported by Länsförsäkringsgruppens Forsknings- & Utvecklingsfond, agreement no. P4/18.

Declarations

Conflict of Interest The author declares no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bajtelsmit, V. L., & VanDerhei, J. L. (1995). Risk aversion and pension investment choices. Wharton Pension Research Council Working Papers. 580. <https://repository.upenn.edu/prc-papers/580>.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Buchanan, J. M. (1972). Rawls on justice as fairness. *Public Choice*, 13(1), 123–128. <https://doi.org/10.1007/BF01718857>.

¹²Whether such orthodox application yields the right result, of course, is a matter of dispute. First, arguing from within the Rawlsian framework, it may fail since the small upside feature may not hold. Second, arguing from without the Rawlsian framework, it may fail for any of the numerous reasons identified by Rawls's critics since 1971. The most famous is probably Nozick (1974, Chapter 7, especially pp. 183–231); for a broader introduction to questions and controversies of distributive justice, see Lamont and Favor (2017).

- Buchanan, J. M., & Tullock, G. (1965). *The calculus of consent: Logical foundations of constitutional democracy*. Ann Arbor: University of Michigan Press.
- Bundy, A. (2017). Smart machines are not a threat to humanity. *Communications of the ACM*, 60(2), 40–42. <https://doi.org/10.1145/2950042>.
- Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5), 82–89. <https://doi.org/10.1145/3376898>.
- Daniels, N. (2020). Reflective equilibrium. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy, summer 2020 edn, Metaphysics Research Lab, Stanford University*.
- Dexe, J., Franke, U., Avatare Nöu, A., & Rad, A. (2020). Towards increased transparency with value sensitive design. In *Artificial Intelligence in HCI. HCI International 2020* (pp. 3–15). Berlin: Springer. https://doi.org/10.1007/978-3-030-50334-5_1.
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. <https://doi.org/10.1126/sciadv.aao5580>.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference Association for Computing Machinery, ITCS '12* (pp. 214–226). New York. <https://doi.org/10.1145/2090236.2090255>.
- Dworkin, R. (1978). Taking rights seriously. Harvard University Press, edition including the appendix “A Reply to Critics”.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center, <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420>.
- Freeman, S. (2019). Original position. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy, Summer 2019 edn, Metaphysics Research Lab, Stanford University*. <https://plato.stanford.edu/archives/sum2019/entries/original-position/>.
- Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6), 16–23. <https://doi.org/10.1145/242485.242493>.
- Galanos, V. (2019). Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. *Technology Analysis & Strategic Management*, 31(4), 421–432. <https://doi.org/10.1080/09537325.2018.1518521>.
- Hamilton, M. (2019). The sexist algorithm. *Behavioral Sciences & the Law*, 37(2), 145–157. <https://doi.org/10.1002/bsl.2406>.
- 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., Ferrari, C., Gummadi, K. P., & Krause, A. (2018). Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 1273–1283). <https://doi.org/10.5555/3326943.3327060>.
- Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). 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). <https://doi.org/10.1145/3287560.3287584>.
- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261–266. <https://doi.org/10.1126/science.aaa8685>.
- Johnson, J. (2019). Artificial intelligence & future warfare: implications for international security. *Defense & Security Analysis*, 35(2), 147–169. <https://doi.org/10.1080/14751798.2019.1600800>.
- Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Fair algorithms for infinite and contextual bandits. arXiv:161009559.
- Keeling, G. (2017). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In *3rd Conference on Philosophy and Theory of Artificial Intelligence* (pp. 259–272). Berlin: Springer. https://doi.org/10.1007/978-3-319-96448-5_29.
- Kitano, H. (2021). Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(29), 1–12. <https://doi.org/10.1038/s41540-021-00189-3>.
- Lamont, J., & Favor, C. (2017). Distributive justice. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy, winter 2017 edn, Metaphysics Research Lab, Stanford University*.
- Lau, J. H., Cohn, T., Baldwin, T., & Hammond, A. (2020). “Deep-speare” crafted Shakespearean verse that few readers could distinguish from the real thing. *IEEE Spectrum*, 57(5), 40–53. <https://doi.org/10.1109/MSPEC.2020.9078455>.
- Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19(2), 107–115. <https://doi.org/10.1007/s10676-017-9419-3>.

- Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, pp. 1–16. <https://doi.org/10.1007/s43681-021-00067-y>.
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>.
- Liang, F. T., Gotham, M., Johnson M., & Shotton, J. (2017). Automatic stylistic composition of bach chorales with deep LSTM. In *The 18th International Society for Music Information Retrieval Conference (ISMIR 2017)* (pp. 449–456).
- 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, 141–163. <https://doi.org/annurev-statistics-042720-125902>.
- Müller, V. C., & Bostrom, N. (2014). Future progress in artificial intelligence: a poll among experts. *AI Matters*, 1(1), 9–11. <https://doi.org/10.1145/2639475.2639478>.
- Nature (2016). More accountability for big-data algorithms. *Nature*, 537(7621), 449. <https://doi.org/10.1038/537449a>.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>.
- Procaccia, A. (2019). AI Researchers Are Pushing Bias Out of Algorithms. <https://www.bloomberg.com/opinion/articles/2019-03-07/ai-researchers-are-pushing-bias-out-of-algorithms>, accessed June 30, 2021.
- Rawls, J. (1999a). *The Law of Peoples*. Harvard: Harvard University Press.
- Rawls, J. (1999b). *A Theory of Justice/Revised Edition*. Oxford: Oxford University Press.
- Resnik, M. D. (1987). *Choices: An Introduction to Decision Theory*. Minnesota: University of Minnesota Press.
- Schmidtz, D. (2006). *Elements of Justice*. Cambridge: Cambridge University Press.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., & et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676), 354–359. <https://doi.org/10.1038/nature24270>.
- Watson, J., & McNaughton, M. (2007). Gender differences in risk aversion and expected retirement benefits. *Financial Analysts Journal*, 63(4), 52–62. <https://doi.org/10.2469/faj.v63.n4.4749>.
- Zhao, Z. Q., Zheng, P., Xu, St., & Wu, X. (2019). Object detection with deep learning: a review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232. <https://doi.org/10.1109/TNNLS.2018.2876865>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.