Philosophy & Technology (2020) 33:225–244
https://doi.org/10.1007/s13347-019-00355-w

RESEARCH ARTICLE

Democratizing Algorithmic Fairness

Pak-Hang Wong[1]

Received: 9 September 2018 / Accepted: 14 May 2019 / Published online: 1 June 2019

Abstract
Machine learning algorithms can now identify patterns and correlations in (big) datasets and predict outcomes based on the identified patterns and correlations. They can then generate decisions in accordance with the outcomes predicted, and decision-making processes can thereby be automated. Algorithms can inherit questionable values from datasets and acquire biases in the course of (machine) learning. While researchers and developers have taken the problem of algorithmic bias seriously, the development of fair algorithms is primarily conceptualized as a *technical* task. In this paper, I discuss the limitations and risks of this view. Since decisions on "fairness measure" and the related techniques for fair algorithms essentially involve choices between *competing* values, "fairness" in algorithmic fairness should be conceptualized first and foremost as a *political* question and be resolved *politically*. In short, this paper aims to foreground the *political* dimension of algorithmic fairness and supplement the current discussion with a deliberative approach to algorithmic fairness based on the accountability for reasonableness framework (AFR).

Keywords Algorithmic bias · Machine learning · Fairness · Democratization · Accountability for reasonableness

1 Introduction

Friedman and Nissenbaum (1996) have shown that computer systems can be biased, that is, computer systems can "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum 1996, p. 332), and the recognition of bias in computer systems has inspired numerous approaches to detect, scrutinize, and avoid bias in computer systems.[1] Despite early

[1]For an overview of major approaches to assess the values embedded in information technology, see Brey (2010).

✉ Pak-Hang Wong
wong@informatik.uni-hamburg.de

[1] Department of Informatics, Universität Hamburg, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany

efforts to combat bias in computer systems, the bias *in* and *through* computing remains today and possibly in a more problematic form. Machine learning algorithms can now identify patterns and correlations in (big) datasets and predict outcomes based on the identified patterns and correlations. They can then generate decisions in accordance with the outcomes predicted, and decision-making processes can thereby be automated. However, algorithms can inherit questionable values from datasets and acquire biases in the course of (machine) learning (Barocas and Selbst 2016; Mittelstadt et al. 2016). Automated algorithmic decision-making also makes it difficult for people to see algorithms as biased either because they, like big data, invoke "the aura of truth, objectivity, and accuracy" (Boyd and Crawford 2012, p. 663) or because they are incomprehensible to an untrained public, and, worse yet, they can even be inscrutable to trained experts (Burrell 2016; Matthias 2004).

The possible harm from algorithmic bias can be enormous as algorithmic decision-making becomes increasingly common in everyday life for high-stakes decisions, e.g., parole decisions, policing, university admission, hiring, insurance, and credit rating. Several high-profile stories in the media have forcibly directed public attention towards the problem of algorithmic bias, and the public has demanded the industry and research community to create "fairer" algorithms.[2] In response, researchers and developers have taken the problem seriously and they have proposed numerous methods and techniques to detect and mitigate bias in algorithms (see, e.g., Lepri et al. 2018; Friedler et al. 2019). However, I shall argue that current responses to algorithmic bias are unsatisfactory, as the development of fair algorithms is primarily conceptualized as a *technical* challenge, where researchers and developers attempt to implement some ideas of "fairness" within algorithms.

In the next section, I explain in more detail what it is to view algorithmic fairness as a technical challenge and illustrate its limitations and risks. I then elaborate the impossibility theorem about algorithmic fairness and the inherent trade-off between fairness and performance in algorithms and argue that they call for an opening-up of the idea of "fairness" in algorithmic fairness, which is *political* in nature. Since decisions on "fairness measure" and the related techniques for fair algorithms essentially involve choices between *competing* values, "fairness" in algorithmic fairness should be conceptualized first and foremost as a *political* question and be resolved *politically*. I suggest that one promising way forward is through democratic communication. If my characterization of the problem of algorithmic fairness is correct, then the task will *not* merely be optimizing algorithms to satisfy some fairness measures and improving relevant techniques for fair algorithms but to consider and accommodate diverse, conflicting interests in a society. The aim of this paper, therefore, is to foreground the *political* dimension of algorithmic fairness and supplement the current discussion with a deliberative approach to algorithmic fairness based on the accountability for reasonableness framework (AFR) developed by Daniels and Sabin (1997, 2008).

[2] The media have reported many cases of (potential) harm from algorithmic decision-making, but the racial bias in the COMPAS recidivism algorithm reported by ProPublica (Angwin et al. 2016; Angwin and Larson 2016), along with Northpointe's (now renamed to "equivant") response to ProPublica's report (Dieterich et al. 2016), have arguably generated the most discussion. The COMPAS recidivism algorithm has since become the paradigmatic case for research on algorithmic bias, with various research citing it as their motivation or using it as a benchmark. Also, see O'Neil (2016) for an accessible discussion of other cases of algorithmic bias.

2 Algorithmic Fairness Is Not Only a Technical Challenge

A recent survey of measures for measuring fairness and discrimination in algorithms describes the task of algorithmic fairness as "translat [ing non-discrimination] regulations mathematically into non-discrimination constraints, and develop[ing] predictive modeling algorithms that would be able to take into account those constraints, and at the same time be as accurate as possible" (Žliobaitė 2017, p. 1061). The quote describes algorithmic fairness primarily as a *technical* challenge to ensure the outcome of an algorithm to approximate the outcome as required by some fairness criteria, while at the same time maintaining its performance with "better" programmed algorithms and "better" pre-processing or post-processing techniques.[3] It requires researchers and developers to first presume some ideas of "fairness" as a benchmark for their works, e.g., the definition of fairness in non-discrimination regulations, for without accepting some ideas of fairness, it is unclear what researchers and developers are programming into an algorithm and what normative standards they are using to assess whether an algorithm is fair or not.[4]

From a technical point of view, an agreement on an appropriate understanding of fairness to be programmed into an algorithm is essential to achieve algorithmic fairness. It is essential because an inappropriate (e.g., "false" or "incorrect") understanding of fairness could defeat the result, as an algorithm cannot be fair insofar as the standard it is based on is *not* fair. Similarly, one can dispute whether an algorithm is fair by questioning the idea of fairness underlying the "fair" algorithm in question. For example, the disagreement between ProPublica and Northpointe (now equivant) over whether the COMPAS recidivism algorithm exhibits racial bias can be attributed to their different understandings of fairness or, more precisely, their understandings of a violation of fairness, namely disparate treatment, where protected features are explicitly used in decision-making, and disparate impact, where the result of a decision disproportionately impacts the protected groups. Northpointe argues that the algorithm is *not biased* because the reoffending rate is roughly the same at each COMPAS scale regardless of a defendant's race; thus, the risk score *means* the same for different races (Dieterich et al. 2016), whereas ProPublica points out that for those who did not reoffend, black defendants are more likely to be classified as having medium or high risk of reoffending than white defendants. Thus, the algorithm is *biased* because one group, i.e., black defendants, is systematically subjected to more severe treatment due to the algorithm's misprediction (Angwin et al. 2016; Angwin and Larson 2016). In this debate, Northpointe and ProPublica have referred to different understandings of the violation of fairness (Corbett-Davies et al. 2016). If there is an agreement on what "fairness" stands for, then algorithmic fairness is indeed a technical challenge of finding the best way to program such an idea of "fairness" into an algorithm.

[3] For a recent overview of the current approaches to algorithmic fairness and different techniques to achieve algorithmic fairness, see Lepri et al. (2018) and Friedler et al. (2019).

[4] This is *not* to claim that the presumed ideas of fairness are unreasonable or idiosyncratic. In fact, some researchers have explicitly referred to the social or legal understandings of fairness in constructing their fairness measures. Still, it is the researchers' *choice* to rely on a specific understanding of fairness, but not the others, for their fairness measures, and their choice is rarely informed by the public. I shall return to this point in my discussion of the AFR-based framework.

Unfortunately, as the dispute on the COMPAS recidivism algorithm demonstrates, the idea of "fairness" in algorithmic fairness is far from being uncontested.

The idea of "fairness" in algorithmic fairness is in many ways contestable, which present an immediate problem to achieving algorithmic fairness. Firstly, there is a growing number of definitions for what "fairness" in algorithmic fairness amounts to, and it seems unlikely for researchers and developers to settle on *the* definition of fairness anytime soon.[5] Secondly, there is a deep disagreement among different philosophical traditions as to what "fairness" should mean and what it entails normatively (Ryan 2006; Binns 2018a), and the same disagreement exists for the closely related concept of "equality of opportunity" as well (Temkin 2017; Arneson 2018). Here, it is useful to reiterate that the disagreement is about the *values* themselves but not the *means* to achieve them. Hence, the disagreement cannot be resolved merely by creating "better" algorithms or using "better" pre-processing and post-processing techniques, as the *normative* standard for assessing what counts as "better," i.e., the very idea of "fairness," is being the locus of the disagreement. In short, the contestability of "fairness" foregrounds the need to settle the meaning of fairness *alongside*, if not *prior to*, the technical tasks as described by Žliobaitė.[6]

What must be emphasized is that the above discussion does not mean to suggest that the industry and research community are unaware of the contestability of fairness and related concepts (see, e.g., Corbett-Davies et al. 2017; Mitchell and Shadlen 2017; Berk et al. 2018; Narayanan 2018). However, there are few attempts to settle the meaning of "fairness" and address the questions of *what ideas of "fairness" are appropriate for algorithmic decision-making* and *why people should accept them*.[7] Without addressing such questions, their attempts to create fair(er) algorithms remain at best incomplete.

As the technical challenge to create fair algorithms can *only* be completed by starting with *some* understandings of fairness, there is the risk of a closing-down of the critical discussion on the ideas of "fairness" in algorithmic fairness, as questioning the meaning of "fairness" may hold researchers and developers back from completing the technical tasks by questioning the foundation of their works. In this respect, the focus on technical challenges of algorithmic fairness risks to discourage critical

[5] For example, Corbett-Davies et al.'s (2017) analysis of the COMPAS recidivism algorithm refers to three definitions of fairness, i.e., statistical parity, conditional statistical parity, and predictive equality. Berk et al.'s (2018) review of fairness in criminal justice risk assessments refers to six definitions of fairness, i.e., overall accuracy equality, statistical parity, conditional procedure accuracy equality, conditional use accuracy equality, treatment equality, and total fairness. Mitchell and Shadlen's (2017) recent summary includes 19 definitions of fairness, and a recent talk by Arvind Narayanan (2018) has increased the number of definitions to 21.

[6] National or international legislation against discrimination may supply the meaning of fairness to researchers and developers for their design and implementation of algorithms. However, there are two potential shortcomings in grounding the "fairness" in fair algorithms on national and international legislation. Firstly, the capacity of algorithms to identify patterns and correlations may engender new types of discrimination that are not based on common protected features, e.g., races and genders. Accordingly, the existing legislation is likely to be insufficient. Secondly, national and international legislation is often difficult and slow to change. Therefore, the idea of "fairness" in algorithmic fairness is likely to be conservative if it is based on the legislation. Of course, national and international legislation remains important to algorithmic fairness for identifying *common* types of discrimination.

[7] For instance, the reason to opt for a specific definition of fairness is often left unarticulated or implicit in the research, except for a few notable exceptions in which researchers and developers acknowledge or reflect on the normative ground of their choice of definition(s). See, e.g., Dwork et al. (2012) and Lipton et al. (2018).

reflection and opening-up of the definition of fairness for public debate and leads to an elitist approach to algorithmic fairness (Skirpan and Gorelick 2017).

3 The Impossibility Theorem, the Inherent Trade-off, and the Political Nature of Algorithmic Fairness

If algorithmic fairness is not merely a technical challenge, then what kind of challenge is algorithmic fairness? Using the impossibility theorem about algorithmic fairness and the inherent trade-off between fairness and performance (or accuracy) in algorithms, I show that algorithmic fairness should not be viewed merely as a *technical* challenge but also a *political* challenge.

Recent research has demonstrated that it is *mathematically* impossible for an algorithm to simultaneously satisfy different popular fairness measures, e.g., disparate treatment and disparate impact, the two fairness measures in the debate on racial bias of the COMPAS recidivism algorithms held by ProPublica and Northpointe, respectively (see, e.g., Friedler et al. 2016; Miconi 2017, Chouldechova 2017; Kleinberg et al. 2017; Berk et al. 2018).[8] The impossibility to simultaneously satisfy two (or more) formalized definitions of fairness suggests that no matter how many definitions of fairness we can arrive at, they will remain contestable by some other definitions of fairness. As Friedler et al. point out, the impossibility theorem is "discouraging if one hoped for a universal notion of fairness" (Friedler et al. 2016, p. 14). The impossibility theorem will also be discouraging to achieving algorithmic fairness from a technical point of view, as no (set of) definition can coherently capture different concerns about fairness at the same time. The immediate lesson from the impossibility theorem is that we need to be more sensitive to the contentious nature of the definition of fairness in the discussion on algorithmic fairness.[9]

In addition to the impossibility theorem, others have pointed to the inherent trade-off between fairness and performance in algorithms (see, e.g., Corbett-Davies et al. 2017; Berk et al. 2018). The trade-off entails that prioritizing fairness in an algorithm will undermine its performance and vice versa. If the algorithm is intended to promote some social goods, and assuming that when functioning well it can achieve this goal, prioritizing fairness necessarily means a loss in those social goods and thus can be conceived as a cost to the society. For instance, Corbett-Davies et al. (2017) have interpreted the trade-off between fairness and performance in the case of the COMPAS recidivism algorithm as a trade-off between fairness (in terms of disparate impact) and public safety, where optimizing for fairness measures is translated as a failure to detain the medium- to high-risk defendants who are more likely to commit violent crimes and thereby threatening public safety.

[8] It is not entirely accurate to describe the incompatibility among different definitions of fairness as "the impossibility theorem." There are indeed situations where some of the definitions of fairness in question can be satisfied simultaneously, but these situations are highly unrealistic, e.g., when we have perfect predictor or trivial predictor that is either always-positive or always-negative (Miconi 2017).

[9] This is not intended to be a knock-down argument against viewing algorithmic fairness primarily as a technical challenge. However, as I have argued the focus on technical tasks can lead to a less critical attitude towards one's idea of "fairness," it is more likely that researchers and developers who see algorithmic fairness primarily as a technical challenge are less sensitive to the contentious nature of the definition of fairness.

For those who value public safety, fairness measures that significantly reduce public safety will be unacceptable. Moreover, they may argue that fairness measures cannot be *genuinely* fair when these measures reduce *their* (public) safety, as optimizing for fairness imposes a risk—or, more precisely, a risk of harm—on people for the benefit of defendants.[10] In other words, prioritizing fairness could unfairly put some members of the public at the risk of harm from violent crimes.[11] Note that this line of argument can be generalized to other algorithms so long as they are designed and implemented to promote social goods. The inherent trade-off between fairness and performance points to the fact that whether the choice of a fairness measure will be considered as acceptable depends on factors that go *beyond* the consideration of fairness as narrowly defined in formalized terms, and it will require balancing fairness with other social goods in the process.[12]

The impossibility theorem and the inherent trade-off between fairness and performance, therefore, raise the following questions: if researchers and developers cannot simultaneously satisfy two (or more) justified understandings of fairness in an algorithm and, at the same time, they have to balance fairness with other social goods, (i) *what should they decide on the definition of fairness, the balance between fairness and social goods, etc. for an algorithm*? And, more importantly, (ii) *how can they justify their decisions to those who will be affected by the algorithm?*

To answer these questions, Narayanan (2018) helpfully reminds us that the different fairness measures can be understood as representing the interests of different stakeholders affected by the algorithm. For example, in the case of the COMPAS recidivism algorithm, judges and parole officers will focus on the (positive) predictive value of the algorithm, i.e., how many correct instances of recidivism can the algorithm identify successfully, and they will also want to ensure irrelevant and sensitive features, such as the defendants' race, do not directly affect the prediction; whereas for the defendants, especially those who are in the protected (minority) group, their concerns will be about the chance of being mistakenly identified by the algorithm as medium- or high-risk, and thereby facing more severe penalty due to the algorithm's error, and this group of individuals will demand the chance of being misclassified not to be significantly greater than the other groups (Narayanan 2018).

Making explicit the relation between stakeholders' interests and fairness measures is important because it invites us to go beyond seeing algorithmic fairness merely as a set of technical challenges to be addressed by programming *some* ideas of "fairness" into an algorithm. The choice of *any* fairness measure will inevitably favor the interests of some groups of stakeholders over the others and thereby benefiting some while harming the others. So construed, algorithmic fairness is not only about designing

[10] There is an important distinction between *actualized* harm and *risk* of harm to be made in the discussion on the fair distribution of risk, see Hayenhjelm (2012) and Hayenhjelm and Wolff (2012). The debate on risk and distributive justice is out of the scope here, but my argument only relies on the assumption that the distribution of risk and benefit is, in fact, an issue of fairness.

[11] Here, the claim about unfairness could at least be grounded on (i) a consequentialist perspective and (ii) a rights-based perspective. From the consequentialist perspective, the unfairness is due to a reduction of overall social good, whereas from the rights-based perspective, individuals have prima facie rights not to be exposed to a risk of harm (see Hayenhjelm and Wolff 2012).

[12] In this respect, the increasing number of researchers being more explicit about the values and normative grounds of various definitions of fairness is a welcoming trend in the research on algorithmic fairness (see, e.g., Dwork et al. (2012); Friedler et al. (2016), Berk et al. (2018), Narayanan (2018)).

and implementing algorithms that satisfy some fairness measures but also about *which ideas of* "*fairness*" and *what other values* should be considered and accommodated in an algorithm. This, in turn, poses a significant ethical and political challenge to those who decide which fairness measure(s) and what other values an algorithm is to include.

Moral philosophers have long argued that imposing a significant risk on people without their consent is prima facie wrong and that consent is morally necessary for imposing risks on individuals (MacLean 1982; Teuber 1990). When an algorithm is devised to make high-stakes decisions *about* or *for* individuals, those who are affected by the algorithm can legitimately question whether the choice of a specific fairness measure and the balance between fairness and performance will put them at a significant risk and insist that their consent is necessary for such a choice to be *morally* defensible.[13]

Similarly, political philosophers and political scientists have argued for the importance of the "all-affected principle" in democracy. The all-affected principle states that those who are significantly affected by a decision ought to be included in the decision-making either directly or indirectly (Dahl 1990, p. 49; cf. Whelan 1983). Those who are affected by the choice of fairness measure and the balance between fairness and performance of algorithms, therefore, ought to have a say in the decision-making process. However, this is complicated by the fact that different fairness measures and different balances between fairness and performance may represent *conflicting* interests of different groups of stakeholders, and each of them will see different choices of fairness measure and the balance between fairness and performance as the "right" one. To settle on an understanding of fairness and to strike a balance between fairness and performance, therefore, are a *political* task that requires researchers and developers to consider and accommodate diverse, conflicting interests of those who are affected by their algorithm. In line with the all-affected principle, it is also a task that should be undertaken by the researchers and developers together *with* the people. Without justifying their decisions on the definition of fairness, fairness measure, and balance between fairness and other social goods to those who will be affected by the algorithm, the people can legitimately question whether the "fair" algorithm is fair *for* them and reject them as truly "fair." In short, the impossibility theorem and the inherent trade-off between fairness and performance call for an opening-up of the definition of fairness and the balance between fairness and performance for public discussion and decision-making.

4 An Accountability for Reasonableness Framework for Algorithmic Fairness

Setting aside the questions of whether and which uses of algorithms in high-stakes decision-making are morally permissible, if we think that algorithmic decision-making can be used in some of those contexts, creating fair(er) algorithms remains an imperative.[14] The impossibility theorem and the inherent trade-off between fairness and performance,

[13] Hansson (2006) has forcibly questioned the applicability of (informed) consent in non-individualistic contexts. Here, the discussion is by no means an argument for the role of (informed) consent in justifying the imposition of risk by algorithms, but it is merely an example of the kind of ethical issues that may arise.

[14] If one considers *every* use of algorithmic decision-making to be morally impermissible, then concerns over fairness in algorithms will cease to exist. The project of achieving fair algorithms presupposes some uses of algorithms to be morally permissible.

however, demonstrate that it is insufficient to view algorithmic fairness merely as a set of technical tasks and show that the chosen fairness measure and the balance between fairness and performance ought to be justified to those who are affected by them. In this respect, achieving algorithmic fairness requires the industry and research community to engage with those who are affected by their algorithms, which is a *political* task. In the remainder of this paper, based on Daniels and Sabin's (1997, 2008) AFR, I outline a framework for algorithmic fairness that accounts for its *political* dimension.

Two caveats must be mentioned before elaborating my AFR-based framework for algorithmic fairness. First, I shall take for granted that there are intractable disagreements among different groups of stakeholders over the question on which conceptions of fairness, fairness measures, and the balance between fairness and performance are the *right* one because stakeholders have diverse, conflicting interests. If there is an uncontroversial agreement on the definition of fairness and of the priority between fairness and other societal goods to be programmed into the algorithm in question, then it may be sufficient to view achieving algorithmic fairness as a technical task.[15] Second, I shall also assume that the interests expressed by different groups of stakeholders and in particular their preferences for specific idea of "fairness," fairness measure, and the balance between fairness and performance are morally and politically justifiable.[16] These two assumptions do not only reiterate the difficulty to conceptualize algorithmic fairness merely as a technical challenge. They also highlight a peculiar condition of liberal democratic society, that is, it is characterized by the *pervasiveness of reasonable disagreement*, or, as Rawls (1993) calls it, *the fact of reasonable pluralism*.[17] Since reasonable disagreement is ineliminable in a liberal democratic society, we can only aim at *reducing disagreement* and *accommodating difference*.[18] It is against this background I introduce Daniels and Sabin's AFR to the problem of algorithmic fairness. As developed by Daniels and Sabin, AFR aims to enable decision-making in the face of pervasive reasonable disagreement, which is also characteristic of the decision-making for fairness measure and the balance between fairness and performance in creating fair algorithms.

Daniels and Sabin's AFR is a response to the problem of limit-setting (and, priority-setting) in the healthcare context.[19] They argue that healthcare is a fundamental human

[15] However, even if there is *no* disagreement among different groups of stakeholders, I take it that the AFR-inspired framework I outline can *enhance* the "fairness" of the decision.

[16] My discussion *only* requires there to be at least *some* choices that are equally justifiable and thereby leading to the requirement for justifying one justifiable choice over another *equally* justifiable choice.

[17] For Rawls, the fact of reasonable pluralism amount to "a pluralism of comprehensive religious, philosophical, and moral doctrines […] a pluralism of incompatible yet reasonable comprehensive doctrines" (Rawls 1993, p. xvi).

[18] Rawls argues that despite there are differences in reasonable comprehensive doctrines, individuals in the society could still achieve mutual agreement on a political conception of justice through overlapping consensus, that is, individuals subscribe to different comprehensive doctrines *can* agree on the political conception of justice with their own reasons and from their own moral points of view (cf. Rawls 1993, p. 134). Yet, the agreement on the political conception of justice is necessarily thin, and thus, it is insufficient to supply fine-grained normative principles to settle substantive value-related issues, e.g., prioritizing the interests of different groups of stakeholders (cf. Daniels 1993).

[19] Daniels and Sabin first proposed AFR in Daniels and Sabin (1997), and Daniels has since defended and applied AFR on various healthcare issues with Sabin and other colleagues. Note that this paper is not an exposition of AFR, and I shall not attempt to survey the extensive discussion on AFR. My discussion of AFR refers primarily to Daniels and Sabin (2008), which incorporate the earlier works on AFR and present the most systematic account of it. However, I shall also refer to earlier works on AFR when I consider them to be more relevant on a specific point under discussion.

good, and people in the society have *reasonable* claims over it. However, even the wealthiest countries will not have enough resources to simultaneously satisfy the claims to different healthcare goods of all, as well as their claims to other fundamental human goods, e.g., education and job opportunities. Any sensible distribution of healthcare goods in a society has to set some limits to the provision of healthcare and to prioritize some claims over the others (Daniels and Sabin 2008, pp. 13–24). Moreover, they also argue that there is neither consensus among people on the how the limits (or priorities) are to be set nor are there fine-grained, substantive normative principles available to arbitrate between *reasonable* claims over different healthcare goods (and other human goods) and between the claims of these human goods from some groups over the others in a democratic society (Daniels and Sabin 2008, pp. 25–41).

The lack of consensus and fine-grained, substantive normative principle suggests that the problem of limit-setting has to be framed as a question of procedural justice, that is, to establish a "process or procedure that most can accept as fair to those who are affected by such decisions. That fair process then determines for us what counts as fair outcome" (Daniels and Sabin 2008, p. 4), because no pre-agreed or universally accepted normative standard can be invoked to justify the limit on healthcare goods (or on other human goods). For Daniels and Sabin, the *normative* question of limit-setting thus has to be reformulated as a question of legitimacy, i.e., "Why or when […] should a patient or clinician who thinks an uncovered service is appropriate […] accept as legitimate the limit setting decision of a health plan or district authority?" (Daniels and Sabin 2008, p. 26) and of *fairness*, i.e., "When does a patient or clinician who thinks an uncovered service appropriate […] have sufficient reason to accept as fair the limit-setting decisions of a health plan or public authority?" (Daniels and Sabin 2008, p. 26). The shift towards procedural justice, i.e., to identify the conditions where decisions are morally and politically acceptable on the ground of legitimacy and reasonableness, allows us to proceed with limit-setting in the absence of a consensus for a universally accepted normative standard. It also highlights Daniels and Sabin's commitment to the democratic ideal of the all-affected principle.

It is useful to elaborate the parallels between the problem of limit-setting in the healthcare context and the problem of algorithmic fairness, as their similarities help to further demonstrate why AFR is suitable in the context of algorithmic fairness. First, the problem of limit-setting and the problem of algorithmic fairness share the issue raised by reasonable disagreement in liberal democratic societies. AFR eschews the search for a pre-agreed or universally accepted normative standard, as it acknowledges that reasonable disagreement makes such a normative standard unlikely. Second, both problems require decisions to be made despite the impossibility to simultaneously satisfy reasonable claims over different goods and from different groups of people. In the case of (non-)provision of healthcare goods, the problem arises from a society's resources being finite, whereas in the case of algorithmic fairness, it is mathematically impossible to satisfy different fairness measures at the same time. Since decisions about healthcare goods—and, in the case of algorithmic fairness, decisions about fairness measure and the balance between fairness and performance—have to be made, AFR's response is to spell out the conditions to ensure the decisions are morally and politically acceptable to those affected by algorithms through inclusion and accommodation of their views and voices.

Moreover, for Daniels and Sabin (2000; also, see 2008, p. 46), the problem of limit-setting is not only a problem for public agencies but also for *private organizations*,

where achieving legitimacy is more challenging because they are directly accountable to the shareholders and only indirectly to other stakeholders. This is also true in the case of algorithmic fairness when the industry and research community are, in fact, the decision-makers—either directly by designing and implementing an algorithm or indirectly by proposing specific ideas of "fairness," fairness measure, and related debiasing and optimization techniques. Here, pre-determining the meaning of fairness by researchers and developers will be morally and politically problematic, as doing so risks neglecting the views and voices of those who will be affected by algorithms. AFR attempts to overcome this risk by specifying the conditions for decision-making where people's views and voices are accounted for.

It is these similarities between the problem of limit-setting and the problem of algorithmic fairness and the potential of AFR to address the peculiar background condition of liberal democratic societies that make it a suitable framework for the problem of algorithmic fairness. We may even view the choice of fairness measure and the balance between fairness and performance as a problem of limit-setting, i.e., setting the limits for fairness and other social goods to be distributed through algorithms in light of reasonable disagreement in a liberal democratic society.

According to AFR, any decision-making process must satisfy four conditions in order to be legitimate and fair. Since Daniels and Sabin's formulation of these conditions is originally intended for the healthcare context, I have modified the four conditions to make them applicable to the problem of algorithmic fairness[20]:

1. Publicity condition: Decisions that establish priorities in meeting [algorithmic fairness] *and their rationales* must be publicly accessible.
2. Relevance condition: The rationales for priority-setting decisions should aim to provide a *reasonable* explanation of why the priorities selected are thought the best way to progressively realize [the value the algorithm aims to provide] or the best way to meet [claims] of the defined population under reasonable […] constraints. Specifically, a rationale will be "reasonable" if it appeals to evidence, reasons, and principles that are accepted as relevant by ("fair minded") people who are disposed to finding mutually justifiable terms of cooperation. An obvious device for testing the relevance of reasons is to include a broad range of stakeholders affected by these decisions so that the deliberation considers the full range of considerations people think are relevant to setting priorities.
3. Revision and appeals condition: There must be mechanisms for challenge and dispute resolution regarding priority-setting decisions and, more broadly, opportunities for revision and improvement of policies in light of new evidence or arguments.
4. Regulative condition: There is public regulation of the process to ensure that conditions (1)–(3) are met. (Daniels 2010, pp. 144–145; original emphasis).

Daniels and Sabin argue that the Publicity condition in AFR ensures the transparency of decisions and decision-making processes, and it allows the public to observe whether

[20] The formulation of the four conditions I quoted is slightly different from the one presented in Daniels and Sabin (2008, p. 45). I refer to this formulation because it is explicitly targeted at the problem of priority-setting, and, as I point out, the choice of fairness measure and balance between fairness and accuracy can be viewed as a priority-setting problem.

the decision-makers are coherent and consistent in their decision-making (Daniels and Sabin 2008, p. 12, pp. 46–47). Making public the reasons for decisions can also force the decision-makers to clarify their rationales and relate them to the people. As reasons become open to public scrutiny, they can contribute to improving the quality of public deliberation and facilitate social learning (Daniels and Sabin 2008, pp. 47–49). Decision-making processes that meet the Publicity condition also show the decision-makers as *principled* and *responsive to the people*, in particular to those who are affected by their decisions and thereby making their decisions legitimate.

Using the COMPAS recidivism algorithm as an illustration, the Publicity condition requires publicizing the choice of fairness measure and the rationales for adopting such a choice, which is indeed what Northpointe did after being criticized by ProPublica (Dieterich et al. 2016; also, see Chouldechova 2017). While it is unfortunate that only after being criticized by ProPublica did Northpointe publicly disclose the fairness measure in the COMPAS recidivism algorithm and its rationales, and that the *right* fairness measure to be used in criminal risk assessment algorithms remains undetermined, it is reasonable to assert that Northpointe, by publicizing its fairness measure and its rationales, does contribute significantly to the social learning of the issue of fairness in criminal risk assessment algorithms, and the same should hold for other algorithms too. Hence, the AFR-based framework requires researchers and developers to *voluntarily* disclose their choices of fairness measure and their rationales for the public benefit.

However, biases in algorithms are often difficult to detect due to the complexity and technicality of algorithms, and it is equally difficult for the public to know *how* an algorithm will affect them and to *whom* it will affect. If the purpose of the Publicity condition is to enhance public deliberation and social learning, I shall add that consequences of an algorithm and to which groups the algorithm will affect ought to be made plain to the public in a *non-technical language*, especially because different fairness measures will have different implications to different groups (Chouldechova and G'Sell 2017). It is only with the knowledge about the consequences of an algorithm and its distributional implications can individuals deliberate competently, and it also prevents self-serving interests from shaping the public deliberation by revealing who is set to benefit and harm by an algorithm.[21] To this end, there are a number of recent projects that help to visualize the distributional implications of different fairness measures, e.g., What-If Tool (https://pair-code.github.io/what-if-tool/) and AI Fairness 360 (https://aif360.mybluemix.net/).

My addition of non-technicality of publicity to the Publicity condition entails that if researchers and developers fail to explain in layman's terms the reasons for their choice of fairness measure and the consequences of their algorithms, the implementation of their algorithms (or the fairness measure) should be considered as morally problematic and politically illegitimate.

Daniels and Sabin intend the Relevance condition to distinguish *valid* reasons from *invalid* reasons in limit-setting decisions by whether they are "accepted as relevant [and appropriate] by ('fair minded') people who are disposed to finding mutually justifiable

[21] Veale and Binns (2017) rightly point out that there are practical difficulties for private organizations to explicate the consequences of an algorithm and its distributional implications, for private organizations may not, or even are not, allowed to possess and process relevant data for such endeavors. I think, however, the responses Veale and Binns provided in their paper can resolve the practical difficulties. In this paper, I cannot discuss their responses in detail, but the proposed responses are compatible with the AFR-inspired framework I develop in here.

terms of cooperation" (Daniels 2010, p. 145), but it has been subjected to different criticisms (see, e.g., Friedman 2008; Lauridsen and Lippert-Rasmussen 2009; Ford 2015; Badano 2018). For instance, the Relevance condition has been criticized as being unspecific, e.g., the important notion of "fair mind" people is left undefined in the condition and only explained with an analogy to (fair) footballers accepting the rules of the game because the rules promote the game (Ford 2015; cf. Daniels and Sabin 2008, pp. 44–45). Without an account of "fair-mindedness" or other normative standards to evaluate the validity of reasons, it is unclear what reasons *should be* included and excluded in the public deliberation. Relatedly, it also calls into question AFR's capacity to weigh reasons for and against some decisions (see, e.g., Landwehr 2013; Tsu 2018).

In a recent critique of the Relevance condition, Badano suggests replacing the Relevance condition with the Full Acceptability condition:

> "[The condition] requires that decision-makers strive to ground [priority-setting] decisions in rationales that each reasonable person can accept, where reasonable persons are understood to be those who are themselves committed to decisions that everyone similarly motivated can accept" (Badano 2018, p. 18).

Badano borrows insights from Nagel (1979, 1991) and Scanlon (1982) and argues that the Full Acceptability condition imposes a tight frame of mind on decision-makers and thus constraining the types of reasons to be presented in public deliberation. He suggests that striving for full acceptability in decisions that inevitably create winners and losers will require decision-makers to settle for a choice that is most acceptable to the person to whom the choice is least acceptable and therefore shift the focus to *individuals' claims* and *the strength of their claims* as the basis for the validity of reason in public deliberation (Badano 2018, pp. 11–14).

Of course, the Full Acceptability condition in itself does not *always* resolve *all* competing claims, but the decision-making can be complemented by a voting mechanism that favors the (most) vulnerable (Tsu 2018) or by an assessment of reasons conducted by independent third parties (Syrett 2002; cf. Veale and Binns 2017; McQuillan 2018). In short, the Full Acceptability condition is useful in limiting the types of reasons in the public deliberation and also in directing us to look at whose claims matter. For instance, the Full Acceptability condition requires engaging with the vulnerable, and it also rejects the use of impersonal reasons, e.g., overall efficiency of the society, to override their claims. In the context of algorithmic fairness, the Full Acceptability condition then requires a close examination of the claims of those who are, or will be, negatively affected by the algorithm (see, e.g., Woodruff et al. 2018), and the decision on the fairness measure must be made such that it is most acceptable to the persons to whom it is least acceptable.[22]

[22] It is useful to caution that both Badano's Full Acceptability condition and Daniels and Sabin's Relevance condition risk over-intellectualized public deliberation and thereby excluding views and voices that are not presented in a rational, argumentative form. Similarly, implicit in the Full Acceptability condition, the importance of achieving consensus, which, in turn, can lead to a suppression of differences. In response to the two concerns, it is useful to explore whether Young's (2000) communicative democracy can broaden the inclusion of views and voices by introducing other modes of communication in public deliberation, e.g., greeting, rhetoric, and narrative; and, whether Young's ideal of differentiated solidarity based on mutual respect and caring but not mutual identification can avoid the suppression of differences (Young 2000, pp. 221–228).

To use the COMPAS recidivism algorithm as an example, the AFR-based framework requires Northpointe to justify its fairness measure and the rationales for it to those who will be affected by the COMPAS recidivism algorithm, e.g., judge, parole officers, and members of the community, and the rationales *must* be ones that are acceptable to those who are (most) vulnerable when the algorithm is in use. Under normal circumstances, I think, the already disadvantaged black defendants and the black community in general are unlikely to accept a fairness measure that will result in a disparate impact on them, no matter how Northpointe tries to justify it. According to the AFR-based framework, Northpointe's choice of fairness measure in the COMPAS recidivism algorithm, therefore, ought to be rejected as morally problematic and politically illegitimate—even if it could be, on other occasions, an appropriate fairness measure. If, however, the recidivism algorithm opts for a different fairness measure with a lower accuracy rate but also a reduced disparate impact, it could be considered by the vulnerable members of the community to be *more* acceptable. Indeed, even if this fairness measure is likely to lead to a *reverse discrimination*, when the community is sufficiently aware of the historical pattern of discrimination against the vulnerable members, those who will be disadvantaged by such a fairness measure, I think, will too view it as acceptable. In reality, of course, the acceptability of fairness measures and their rationales can only be determined in *actual* contexts of uses, the AFR-based framework foregrounds the need to include the (vulnerable) members of community in creating fair algorithms.

As society continues to evolve with new knowledge and technology, publicizing decisions and their rationales and participating in reasonable public deliberation cannot be seen as a one-off exercise. It should be viewed as an ongoing process that responds to new insights and evidence related to the decisions and their consequences as well as new options made possible by research and innovation. Here, ProPublica's critique of the COMPAS recidivism algorithm and the subsequent research that follows from the critique are a good case in point. While Northpointe did respond to the criticisms, there is *no* mechanism from *within* or *outside* Northpointe that sufficiently demonstrates the criticisms have been addressed satisfactorily. For the COMPAS recidivism algorithm to be morally and politically acceptable, Northpointe should have established *meaningful* feedback procedures and systems for its users, targets, and other relevant parties. To the AFR-based framework, it is the lack of such mechanisms that renders Northpointe's fairness measure and the COMPAS recidivism algorithm morally and politically problematic.

The Revision and Appeal condition is necessary to cope with the rapidly changing social and technological environment. In effect, without proper means to review and revise previous decisions, the cost of mistakes will be excessively high and decision-making could be hindered. In addition, good mechanisms for challenge and dispute resolution can strengthen the legitimacy of decision-making and contribute to the social learning of the problem at hand, as they give people, notably those who might not have been included in the initial decision-making, an opportunity to be heard and invite them to reflect on the valid reasons that have been expressed to support the original decisions (Daniels and Sabin 2008, 58–59). The need to review and revise decisions in the context of algorithmic fairness is even more pressing, as the use of algorithm in high-stakes decision-making is still in its early days, and we can readily expect new research to disrupt our preconception of our (original) choice of fairness measures and its consequences.

Finally, the Regulative condition is proposed to ensure private organizations' adoption of the Publicity condition, the Full Acceptability condition (or, the Relevance condition originally proposed by Daniel and Sabin), and the Revision and Appeal condition. In the context of algorithmic fairness, the Regulative condition calls attention to the necessity of regulations and public agencies to enforce the three conditions on public and private organizations that use algorithms for decision-making.

To summarize, the four conditions in AFR specify when a decision is considered to be legitimate and fair even when there is reasonable disagreement but no fine-grained, substantive normative principle to settle such a disagreement. AFR takes seriously the contestable nature of the problem of priority-setting, and it does not presume a "right" answer at the beginning, because people in a liberal democratic society can reasonably disagree with each other about the "right" answer, but AFR sees the answer to emerge from public deliberation. It is the contestability of "fairness" in algorithmic fairness and the political nature of the problem of algorithmic fairness that make AFR particularly suitable to the problem of algorithmic fairness. An AFR-based framework for algorithmic fairness opens up the ideas of "fairness" in algorithmic fairness to the public, especially to those who are affected by the use of algorithm, and it attempts to ground the choice of the definition of fairness, fairness measure, and the balance between fairness and performance with democratic communication. In this respect, the normative foundation of the AFR-based framework rests ultimately on deliberative democracy (Daniels and Sabin 2008, pp. 34–36; also, see Gutmann and Thompson 1996, 2004; Habermas 1996; Young 2000).[23]

It is thus worth to reemphasize that the AFR-based framework I proposed in this section does not aim to offer a different *substantive* idea of fairness for creating fair algorithms, but it shifts the focus to the *processes* or *procedures* for determining which ideas of "fairness" and other societal values to be considered and accommodated in designing and implementing algorithmic decision-making systems: the AFR-based framework requires the industry, research community, and the people to be public about their decisions and their rationales and mandates the choice (and the reasons for it) to be one that is most acceptable to those who are being adversely affected. The key to legitimate and fair decisions and fair algorithms is, therefore, *the exchange of reasons*. Also important is the AFR-based framework's insistence on decision-making about algorithmic fairness should be viewed as an ongoing process that is defeasible as new knowledge and technologies come onto the scene. In short, the decision-making of "fairness" in algorithmic fairness should be open to *new* reasons.

Here, the emphasis on the exchange of reasons and the view that decision-making of "fairness" in algorithmic fairness should be regarded as an ongoing process can be helpfully illustrated by a contrast with Grgić-Hlača et al.'s (2018) recent proposal for a

[23] The more fundamental questions for the AFR-based framework, therefore, are about (i) the normative and practical viability of deliberative democracy and (ii) the proper scope of it. In other words, a more comprehensive account of the AFR-based framework requires one to defend deliberative democracy as a better alternative than other forms of democracy and to work out the institutional arrangements where individuals' views and voices can be adequately communicated. It must also specify whose views and voices are to be included, e.g., citizens vs. non-citizens in the democratic society, and what questions are open for democratic deliberation, e.g., national security issues. Debates on theoretical and practical aspects of deliberative democracy have generated an enormous amount of research that I cannot summarize in this paper, but I shall acknowledge the significant role deliberative democracy in normatively grounding my AFR-based framework. For a review of the prospect of deliberative democracy, see Curato et al. (2017).

technical approach to *procedural justice* for algorithmic fairness. In their work, they crowdsource individuals to vote on the features they, the voters, consider to be fair to include in the algorithmic decision-making system and program the algorithm according to the input from the crowdsourced individuals (Grgić-Hlača et al. 2018). Surely, their technical approach is conscious of the contestable nature of the ideas of "fairness," and it is, in an important sense, *open* to the public, i.e., the fairness measure is determined *by* and *after* people's vote. Yet, their technical approach remains insufficient, at least, in terms of the conditions required by the AFR-based framework, because it is only based on an aggregation of preferences, which does not involve a genuine *exchange* of reasons. For the AFR-based framework, it is the *reason-giving* and *reason-responding* in the exchange of reasons that show respect to individuals' views and voices and recognize their differences. Moreover, it is through the exchange of reasons that different parties involved learn more deeply about the problem of algorithmic fairness and learn from those who are adversely affected by algorithmic decision-making. From this point of view, a mere aggregation of preferences will not suffice.

5 Conclusion

In this paper, I attempt to show that there are more to creating algorithmic fairness than a set of technical tasks, i.e., there is an important political dimension in the problem of algorithmic fairness due to the contentious nature of the ideas of "fairness" and the fact that a decision on fairness measure and the balance between fairness and performance is in effect about *competing* values. One of the main contributions of this paper, therefore, is to explicitly formulate the problem of algorithmic fairness in part as a *political* challenge and to draw attention to the need to resolve it by *political means*.[24] To this end, I have proposed a version of AFR to address the political challenge in creating fair algorithms.

According to the AFR-based framework, an algorithm is considered to be *genuinely* fair and thus morally acceptable and politically legitimate, when the four conditions are satisfied. More specifically, it demands the strengthening of the regulation of algorithmic decision-making systems to ensure their design and implementation to be public, reasonable, and revisable (i.e., the Regulative condition), that is, the AFR-based framework requires the industry and research community to account for the interests of those who are affected by their algorithms through (i) making public and in plain language the ideas of "fairness," fairness measure, and the balance between fairness and other societal values when designing and implementing algorithmic decision-making systems (i.e., the Publicity condition), (ii) grounding their decisions with reasons that are acceptable by those who are most adversely affected (i.e., the Full Acceptability condition), and (iii) establishing suitable mechanisms to resolve disputes

[24] Binns (2018b) is an important exception to this claim, where he explores the phenomenon of algorithmic accountability in terms of the democratic ideal of public reason. While there are affinities between my discussion and Binns' account, there are two important differences. Firstly, I attempt to demonstrate the political dimension in the problem of algorithmic fairness is due to its internal features, particularly the impossibility theorem and the inherent trade-off between fairness and accuracy. Secondly, I attempt to offer a specific approach to ground decision-makers' accountability with Daniels and Sabin's AFR.

arise from the use of their algorithms and being open to adjustment in light of new reasons (i.e., the Revision and Appeals condition).

The requirements of the AFR-based framework are not unique, and similar requirements have been expressed in major ethical and governance principles for algorithm design and implementation (see, e.g., Diakopoulos et al. n.d.; USACM 2017; Reisman et al. 2018; Partnership on AI 2019). For instance, Algorithmic Impact Assessments (AIAs), a detailed framework proposed by AI Now Institute, has included four policy goals:

> "1. Respect the public's right to know which systems impact their lives by publicly listing and describing automated decision systems that significantly affect individuals and communities;
> 2. Increase public agencies' internal expertise and capacity to evaluate the systems […], so that they can anticipate issues that might raise concerns, such as disparate impacts or due process violations;
> 3. Ensure greater accountability of automated decision systems by providing a meaningful and ongoing opportunity for external researchers to review, audit, and assess these systems […]; and
> 4. Ensure that the public has a meaningful opportunity to respond to and, if necessary, dispute the use of a given system or an agency's approach to algorithmic accountability" (Reisman et al. 2018, p. 5).

Similarly, Partnership on AI (2019) has recently published a report on algorithmic criminal risk assessment tools that include ten requirements for the responsible use of these tools. Particularly relevant are the requirements for "Governance, Transparency, and Accountability":

> "Requirement 7: Policymakers must ensure that public policy goals are appropriately reflected in these tools.
> Requirement 8: Tool designs, architectures, and training data must be open to research, review and criticism.
> Requirement 9: Tools must support data retention and reproducibility to enable meaningful contestation and challenges.
> Requirement 10: Jurisdictions must take responsibility for the post-deployment evaluation, monitoring, and auditing of these tools." (Partnership on AI 2019)[25]

The Publicity condition in the AFR-based framework captures policy goal (1) of AIAs and part of Requirement 8 in the report; Requirement 7 in the report can be viewed as jointly derived from the Publicity condition and the Full Acceptability condition. Policy goals (2), (3), and (4) and Requirements 8 and 9 are describing versions of the Revision

[25] The other requirements listed in the report are related to "Accuracy, Validity, and Bias," i.e., "Requirement 1: training datasets must measure the intended variables," "Requirement 2: bias in statistical models must be measured and mitigated," and "Requirement 3: tools must not conflate multiple distinct predictions" and to "Human-Computer Interface Issues," i.e., "Requirement 4: predictions and how they are made must be easily interpretable," "Requirement 5: tools should produce confidence estimates for their predictions," and "Requirement 6: users of risk assessment tools must attend trainings on the nature and limitations of the tools."

and Appeals condition. Finally, Requirement 10 can be understood as a form of the Regulative condition.[26]

The similarities between the AFR-based framework and major ethical and governance principles should not be surprising, as the four conditions in the AFR-based framework succinctly capture the basis of what is it for decision-makers to be accountable and their decisions to be morally acceptable and politically legitimate. The major ethical and governance principles do offer useful guidelines and recommendations to researchers and developers in creating fair algorithms, but the AFR-based framework I proposed provides these ethical and governance principles with a *philosophically coherent* and *normative strong* foundation.

Daniels and Sabin stress that AFR is not intended to be "merely a theoretical, but [a practical] solution to the fairness and legitimacy problems" (Daniels and Sabin 2008, p. 27). In this respect, the AFR-based framework outlined in this paper should also be viewed as a *practical* solution. To further operationalize the AFR-based framework, however, two lines of future research are necessary: firstly, a more detailed analysis of the Full Acceptability condition (or the Relevance condition) is required to spell out the ways to adjudicate between different reasons and, secondly, a careful study of the methods of public deliberation that enable agreement and accommodate differences and their limitations is also required in order for the framework to apply in practice.[27] In short, the goal of my discussion of AFR in this paper is only modest, that is, by foregrounding the similarities between the problem of limit-setting in the healthcare context and the problem of algorithmic fairness, I reformulate the problem of algorithmic fairness as a problem of limit-setting and demonstrate AFR to be a promising framework for the problem.

References

ACM US Public Policy Council [USACM] (2017). *Statement on algorithmic transparency and accountability*. Association for Computing Machinery. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf. Accessed 23 April 2019.

Angwin, J. Larson, J. (2016) ProPublica responds to company's critique of machine bias story. *ProPublica*, July 29, 2016. Available online at: https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story

Angwin, J. Larson, J. Mattu, S. Kirchner, L. (2016) Machine bias. *ProPublica*, May 23, 2016. Available online at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arneson, R. (2018). Four conceptions of equal opportunity. *The Economic Journal, 128*(612), F152–F173. https://doi.org/10.1111/ecoj.12531.

Badano, G. (2018). If you're a Rawlsian, how come you're so close to utilitarianism and intuitionism? A critique of Daniels's accountability for reasonableness. *Health Care Analysis, 26*(1), 1–16.

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

[26] This is not the only possible mapping of the four conditions with the policy goals of AIAs and requirements in the report by Partnership on AI. The aim of this exercise is to demonstrate the affinity of the AFR-based framework with major ethical and governance principles.

[27] For example, see Ney and Verweij (2015) for an excellent discussion of different methods to engage the public and to accommodate the normative principles and values of different, conflicting worldviews in relation to wicked problems, but also see Hagendijk and Irwin (2006) for a discussion about the difficulties for public deliberation and deliberative democracy in science and technology policies.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods and Research, OnlineFirst.* https://doi.org/10.1177/0049124118782533.

Binns, R. (2018a). Fairness in machine learning: lessons from political philosophy. *Journal of Machine Learning Research, 81*, 1–11.

Binns, R. (2018b). Algorithmic accountability and public reason. *Philosophy and Technology, 31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*(5), 662–679.

Brey, P. A. E. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 41–58). Cambridge: Cambridge University Press.

Burrell, J. (2016). How the machine 'thinks:' understanding opacity in machine learning algorithms. *Big Data and Society, 3*, 1–12. https://doi.org/10.1177/2053951715622512.

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047.

Chouldechova, A. G'Sell, M. (2017) Fairer and more accurate, but for whom? Poster presented at: The 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017).

Corbett-Davies, S. Pierson, E. Feller, A. Goel, S. (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Washington Post, October 17, 2016. Available Online at: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Corbett-Davies, S. Pierson, E. Feller, A. Goel, S. Huq, A. (2017). Algorithmic decision making and the cost of fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17), 797–806. https://doi.org/10.1145/3097983.3098095.

Curato, N., Dryzek, J. S., Ercan, S. A., Hendriks, C. M., & Niemeyer, S. (2017). Twelve key findings in deliberative democracy research. *Daedalus, 146*(3), 28–38.

Dahl, R. A. (1990). *After the revolution? Authority in a good society*, Revised Edition. New Haven: Yale University Press.

Daniels, N. (1993). Rationing fairly: programmatic considerations. *Bioethics, 7*(2–3), 224–233.

Daniels, N. (2010). Capabilities, opportunity, and health. In H. Brighouse & I. Robeyns (Eds.), *Measuring justice: primary goods and capabilities* (pp. 131–149). Cambridge: Cambridge University Press.

Daniels, N., & Sabin, J. (1997). Limits to health care: fair procedures, democratic deliberation, and the legitimacy problem for insurers. *Philosophy & Public Affairs Public Affairs, 26*(4), 303–350.

Daniels, N., & Sabin, J. (2000). The ethics of accountability in managed care reform. *Health Affairs, 17*(5), 50–64.

Daniels, N., & Sabin, J. (2008). *Setting limits fairly: Learning to share resources for health* (2nd ed.). New York: Oxford University Press.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (n.d.). *Principles for accountable algorithms and a social impact statement for algorithms*. Fairness, Accountability, and Transparency in Machine Learning. http://www.fatml.org/resources/principles-for-accountable-algorithms. Accessed 23 April 2019.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: demonstrating accuracy equity and predictive parity*. Northpoint Inc. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf. Accessed 23 April 2019.

Dwork, C, Hardt, M., Pitassi, T., Reingold, O., Zemel, R (2012) Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226. https://doi.org/10.1145/2090236.2090255.

Ford, A. (2015). Accountability for reasonableness: the relevance, or not, of exceptionality in resource allocation. *Medicine, Health Care and Philosophy, 18*(2), 217–227.

Friedler, S. Scheidegger, C. Venkatasubramanian, S. (2016) On the (Im)possibility of fairness. arXiv preprint, arXiv:1609.07236.

Friedler, S. A. Scheidegger, C. Venkatasubramanian, S. Choudhary, S. Hamilton, E. P. Roth, D. (2019) A comparative study of fairness-enhancing interventions in machine learning, Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 329–338. https://doi.org/10.1145/3287560.3287589.

Friedman, A. (2008). Beyond accountability for reasonableness. *Bioethics, 22*(2), 101–112.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Grgić-Hlača, N. Zafar, M. B. Gummadi, K. P. Weller, A. (2018) Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning, Proceeding of the thirty-second AAAI conference on artificial intelligence, 51–60.

Gutmann, A., & Thompson, D. (1996). *Democracy and disagreement*. Cambridge: Harvard University Press.

Gutmann, A., & Thompson, D. (2004). *Why deliberative democracy*. Princeton: Princeton University Press.

Habermas, J. (1996). *Between facts and norms: contributions to a discourse theory of law and democracy*. Cambridge, MA: MIT Press.

Hagendijk, R., & Irwin, A. (2006). Public deliberation and governance: engaging with science and technology in contemporary Europe. *Minerva, 44*(2), 167–184.

Hansson, S. O. (2006). Informed consent out of context. *Journal of Business Ethics, 63*(2), 149–154.

Hayenhjelm, M. (2012). What is a fair distribution of risk? In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (Eds.), *Handbook of risk theory: epistemology, decision theory, ethics, and social implications of risk* (pp. 910–929). Dordrecht: Springer.

Hayenhjelm, M., & Wolff, J. (2012). The moral problem of risk impositions: a survey of the literature. *European Journal of Philosophy, 20*(S1), E26–E51.

Kleinberg, J. Mullainathan, S. Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores, Proceedings of 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), 43. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Landwehr, C. (2013). Procedural justice and democratic institutional design in health-care priority-setting. *Contemporary Political Theory, 12*, 296–317.

Lauridsen, S., & Lippert-Rasmussen, K. (2009). Legitimate allocation of public healthcare: beyond accountability for reasonableness. *Public Health Ethics, 2*(1), 59–69.

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

Lipton, Z., Chouldechova, A., & McAuley, J. (2018). Does mitigating ML's impact disparity require treatment disparity? In *Proceedings of the Neural Information Processing Systems Conference 2018 (NIPS 2018)*. http://papers.nips.cc/book/advances-in-neural-information-processing-systems-31-2018 . Accessed 23 April 2019.

MacLean, D. (1982). Risk and consent: philosophical issues for centralized decisions. *Risk Analysis, 2*(2), 59–67.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

McQuillan, D. (2018). People's councils for ethical machine learning. *Social Media + Society*, 1–10. https://doi.org/10.1177/2056305118768303.

Miconi, T. (2017) The impossibility of "fairness": a generalized impossibility result for decisions. arXiv preprint, arXiv:1707.01195.

Mitchell, S., & Shadlen, J. (2017). Fairness: notation, definitions, data, legality. https://shiraamitchell.github.io/fairness/old.html. Accessed 2 May 2019.

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

Nagel, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

Nagel, T. (1991). *Equality and partiality*. Oxford: Oxford University Press.

Narayanan, A. (2018). 21 fairness definitions and their politics. https://www.youtube.com/watch?v=jIXIuYdnyyk. Accessed 23 April 2019.

Ney, S., & Verweij, M. (2015). Messy institutions for wicked problems: how to generate clumsy solutions? *Environment and Planning C: Politics and Space, 33*(6), 1679–1696.

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

Partnership on AI (2019). *Report on algorithmic risk assessment tools in the US criminal justice system*. Partnership on AI. https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/. Accessed 2 May 2019.

Rawls, J. (1993). *Political liberalism*. New York: Columbia University Press.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf. Accessed 23 April 2019.

Ryan, A. (2006). Fairness and philosophy. *Social Research, 73*(2), 597–606.

Scanlon, T. (1982). Contractualism and utilitarianism. In A. Sen & B. Williams (Eds.), *Utilitarianism and beyond* (pp. 103–128). Cambridge: Cambridge University Press.

Skirpan, M. Gorelick, M. (2017) The authority of "fair" in machine learning. Paper presented at: The 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017). arXiv: 1706.09976.

Syrett, K. (2002). Nice work? Rationing, review and the 'legitimacy problem' in the new NHS. *Medical Law Review, 10*(1), 1–27.

Temkin, L. (2017). The many faces of equal opportunity. *Theory and Research in Education, 14*(3), 255–276. https://doi.org/10.1177/1477878516680410.

Teuber, A. (1990). Justifying risk. *Daedalus, 119*(4), 235–254.

Tsu, P. S.-H. (2018). Can the AFR approach stand up to the test of reasonable pluralism? *The American Journal of Bioethics, 18*(3), 61–62. https://doi.org/10.1080/15265161.2017.1418929.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. *Big Data and Society, 2017*, 1–17. https://doi.org/10.1177/2053951717743530.

Whelan, F. G. (1983). Democratic theory and the boundary problem. In J. R. Pennock & J. W. Chapman (Eds.), *Liberal democracy* (pp. 13–47). New York: New York University Press.

Woodruff, A. Fox, S. E. Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Paper No. 656*. https://doi.org/10.1145/3173574.3174230.

Young, I. M. (2000). *Inclusion and democracy.* New York: Oxford University Press.

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