Democratizing Algorithmic Fairness

Pak-Hang Wong
Department of Informatics, Universität Hamburg

EMAIL: wong@informatik.un-hamburg.de

ADDRESS: Universität Hamburg,
Department of Informatics,
Vogt-Kölln-Straße 30,
22527 Hamburg,
GERMANY

ABSTRACT

Algorithms can now identify patterns and correlations in the (big) datasets, and predict outcomes based on those identified patterns and correlations with the use of machine learning techniques and big data, decisions can then be made by algorithms themselves in accordance with the predicted outcomes. Yet, algorithms can inherit questionable values from the datasets and acquire biases in the course of (machine) learning, and automated algorithmic decision-making makes it more difficult for people to see algorithms as biased. While researchers have taken the problem of algorithmic bias seriously, but the current discussion on algorithmic fairness tends to conceptualize 'fairness' in algorithmic fairness primarily as a *technical* issue and attempts to implement pre-existing ideas of 'fairness' into algorithms. In this paper, I show that such a view of algorithmic fairness as technical issue is unsatisfactory for the type of problem algorithmic fairness presents. Since decisions on fairness measure and the related techniques for algorithms essentially involve choices between *competing* values, 'fairness' in algorithmic fairness should be conceptualized first and foremost as a *political* issue, and it should be (re)solved by democratic communication. The aim of this paper, therefore, is to explicitly reconceptualize algorithmic fairness as a *political* question and suggest the current discussion of algorithmic fairness can be strengthened by adopting the accountability for reasonableness framework.

KEYWORDS: *Algorithmic Bias, Machine Learning, Fairness, Democratization, Accountability for Reasonableness*.

ACKNOWLEDGEMENT: -

Democratizing Algorithmic Fairness

Introduction

Batya Friedman and Helen Nissenbaum (1996) have long 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 & Nissenbaum 1996, 332), and the recognition of bias in computer systems has inspired numerous approaches to detect, scrutinize and prevent biases in computer systems.[1] Despite the early efforts to combat bias, bias *in* and *through* computing remains today, and in a possibly more problematic guise. Particularly, algorithms can now identify patterns and correlations in the (big) datasets, and predict outcomes based on those identified patterns and correlations with the use of machine learning techniques and big data, and decisions can then be made by algorithms themselves in accordance with the predicted outcomes. In other words, decision-making processes can be completely automated. Yet, algorithms can inherit questionable values from the datasets and acquire biases in the course of (machine) learning (Barocas & Selbst 2016; Mittelstadt *et al.* 2016). Automated algorithmic decision-making, however, makes it more difficult for people to see algorithms as biased either because they, like big data, invoke "the aura of truth, objectivity, and accuracy" (boyd & Crawford 2012, 663), or because they are incomprehensible to an untrained public—and, worst still, they can even be inscrutable to the trained experts (Burrell 2016; Matthias 2004).

The possible harm from algorithmic bias can be enormous as algorithmic decision-making become increasingly common in everyday life for high-stake decisions, e.g. parole decisions, policing, university admission, hiring, insurance and credit rating, etc. Indeed, a number of high-profiles stories in the media have forcibly directed public attention towards the problem of algorithmic bias, and a heightened pressure is mounted on both the industry and research community to create 'fairer' algorithms.[2] In response to algorithmic bias, researchers have taken the problem seriously and numerous responses have been proposed for detecting and mitigating bias in algorithms (see, e.g. Lepri *et al*. 2017; Friedler *et al.* 2018). Yet, I shall argue that the current responses to algorithmic bias is unsatisfactory, as 'fairness' in algorithmic fairness is primarily conceptualized as a *technical* issue and the major task of researchers is taken to be implementing some pre-existing ideas of 'fairness' into algorithms.

In the next section, I explain in more detail what it is to conceptualize algorithmic fairness as technical issue and discuss why this view is unfit for the type of problem algorithmic fairness

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

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

intends to answer. I then elaborate the impossibility theorem about algorithmic fairness and the inherent trade-off between fairness and accuracy (or performance) in algorithms and argue that they call for an opening-up of the idea of 'fairness' in algorithmic fairness. Since decisions on fairness measure and the related techniques for algorithms essentially involve choices between *competing* values, 'fairness' in algorithmic fairness should be conceptualized first and foremost as a *political* issue and to be (re)solved *politically*. I suggest that one promising way forward is through democratic communication. In short, if my characterization of algorithmic fairness is correct, its task will *not* simply be optimizing algorithms to satisfy some fairness measures and polishing relevant techniques for algorithms, but the task is to consider and accommodate the diverse, conflicting interests in the society. The aim of this paper, therefore, is to explicitly reconceptualize algorithmic fairness as a *political* question and to supplement the current discussion on algorithmic fairness with a deliberative approach to algorithmic fairness based on the accountability for reasonableness (AFR) framework (Daniels & Sabin 1997, 2008).

Algorithmic Fairness as a Technical Issue

A recent survey of measures for measuring fairness and discrimination in algorithms described 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, 1061).[3] So construed, algorithmic fairness is primarily conceptualized as a *technical* task of creating 'better' algorithms and using 'better' pre-processing or post-processing techniques to enable the outcome of an algorithm to approximate the outcome as specified by some fairness measures while at the same time maintaining its performance. I shall label this view of algorithmic fairness as *the technical view of algorithmic fairness*. Note that to the technical view algorithmic fairness as I have described requires researchers to presume some pre-existing ideas of 'fairness', e.g. the definitions of fairness in non-discrimination regulations, as a benchmark, for without them it is unclear what researchers are implementing into an algorithm and what normative standards they are using to assess whether an algorithm is fair.[4]

Essential to the technical view of algorithmic fairness is a pre-agreement over an appropriate understanding of fairness to be implemented into algorithms. This pre-agreement can be based on national or international legislations against discrimination,

[3] For recent overviews of current approaches and techniques to algorithmic fairness, see Lepri *et al.* (2017), Friedler *et al.* (2018). It should be pointed out that the reason to opt for particular definitions of fairness is often left implicit or unarticulated in the research, but there are some notable exceptions where researchers explicitly acknowledge or reflect on the normative grounds for their choice of definitions, see, e.g. Dwork *et al.* (2012), Lipton *et al.* (2018).

[4] This is *not* to claim that the pre-existing ideas of fairness are unreasonable or idiosyncratic. In fact, some researchers have explicitly referred to social or legal understandings of fairness in constructing their fairness measures. Nonetheless, it is researchers' *choice* to use some understandings of fairness over 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 deliberative approach.

researchers' consensus, etc., and then the fair algorithms are to be derived from it. It is essential because an inappropriate (e.g. 'false' or 'incorrect') understanding of fairness could derail the subsequent results because the fair algorithms will be based on insufficient *normative* grounds. Similarly, the 'fairness' of a fair algorithm could be challenged, if one could dispute a specific understanding 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, i.e. disparate treatment and disparate impact. Northpointe argued that the algorithm is *not biased* because the reoffending rate is roughly the same at each COMPAS scale regardless of the defendant's race, thus the risk score *means* the same for different races (Dieterich *et al.* 2016), whereas ProPublica pointed out that for those who did not reoffend, blacks are more likely to be classified as having medium or high risk of reoffending than whites, thus the algorithm is *biased* because one group, i.e. blacks, is systematically subjected to harsher treatment due to the algorithm's misprediction (Angwin *et al.* 2016; Angwin & Larson 2016). Here, Northpointe and ProPublica refer to different understandings of fairness in the debate (Corbett-Davies *et al.* 2016). It seems, therefore, that technical view of algorithmic fairness requires an *uncontroversial* understanding of fairness.

Indeed, if there is an agreement on what 'fairness' stands for, then the task of algorithmic fairness will be to find the best ways to operationalize such an idea of 'fairness' in algorithms, and algorithmic fairness can certainly be seen as a technical issue. Unfortunately, as the example of the COMPAS recidivism algorithm has demonstrated, the idea of 'fairness' is far from being uncontestable.

The idea of 'fairness' in algorithmic fairness is in many ways contestable, thus presents an immediate challenge to the technical view of algorithmic fairness. Firstly, there is a growing number of definitions for what 'fairness' amounts to in algorithmic fairness, and it seems unlikely for researchers to settle on *the* definition of fairness anytime soon.[5] Secondly, there is deep disagreement among different philosophical traditions as to what the concept of 'fairness' should capture and what does it imply normatively (Ryan 2006; Binns 2018). The same disagreement also exists for the closely related concept of 'equality of opportunity' as well (Temkin 2017; Arneson 2018). Now, 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 simply by creating 'better' algorithms or using 'better' techniques, as the *normative* standard for assessing what counts as 'better', i.e. the very idea of 'fairness', is being the locus of disagreement. When there is no uncontroversial understanding of fairness

[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); and, Berk *et al.*'s (2017) review of fairness in criminal justice risk assessments refers to *six* definitions (i.e. overall accuracy equality, statistical parity, conditional procedure accuracy equality, conditional use accuracy equality, treatment equality, total fairness). Mitchell and Shadlen's (2017) recent summary of definitions of fairness in research on algorithmic fairness includes *nineteen* definitions, and a recent talk by Arvind Narayanan (2018) has increased the number of definitions to *twenty-one*.

researchers can depend on, the technical view of algorithmic fairness appears to be untenable.[6]

More importantly, the technical view of algorithmic fairness encourages a closing-down of algorithmic fairness, and not to question the ideas of 'fairness' in algorithmic fairness, as doing so will divert researchers from the task of implementation. In this respect, the technical view of algorithmic fairness discourages critical reflection on the ideas of 'fairness' in the discussion on algorithmic fairness and avoids the opening-up of the definition of fairness for public debate, and we may view this as elitist (Skirpan & Gorelick 2017).

The Impossibility Theorem and the Inherent Trade-off in Algorithmic Fairness

The impossibility theorem about algorithmic fairness and the inherent trade-off between fairness and accuracy (or performance) in algorithms further demonstrate why algorithmic fairness should not be viewed merely as a *technical* issue but a *political* issue.

It has been demonstrated by a number of researchers that it is *mathematically* impossible for an algorithm to simultaneously satisfy 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; Kleinberg *et al.* 2016; Chouldechova 2017; Berk *et al.* 2017; Miconi 2017).[7] The impossibility to simultaneously satisfy two (or more) formalized definitions of fairness means 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.* nicely point out, the impossibility theorem is "discouraging if one hoped for a universal notion of fairness" (Friedler *et al.* 2016, 14). The impossibility theorem will also be discouraging to the technical view of algorithmic fairness, as no (set of) definition can coherently capture different concerns about fairness at the same time, thus making the prospect of the uncontroversial understanding of fairness unlikely. Here, the lesson from the impossibility theorem is that we need to be more sensitive to the contentious nature of the definitions of fairness in algorithmic fairness.[8]

[6] Here, national or international legislations against discrimination may offer *the* normative standards for the researchers in designing and implementing algorithms. There are, however, two problems in grounding 'fairness' in algorithmic fairness on national and international legislations. Firstly, algorithms' capacity to identify patterns and correlations made possible new types of discrimination that are *not* based on common protected features, e.g. races, genders, etc. Hence, algorithmic fairness that is based on existing legislations are likely to be insufficient. Secondly, changes in national and international legislations are often difficult and slow. Algorithms fairness is likely to be conservative if it is based on the legislations. This is, however, not to argue that national and international legislations are unimportant, they remain useful to identify common types of discrimination.

[7] It is not entirely accurate to describe the incompatibility between different definitions of fairness as 'the impossibility theorem', as there are indeed situations where the definitions of fairness in question can be satisfied simultaneously. However, those situations are highly unrealistic, e.g. when we have perfect predictor or trivial (i.e. always-positive or always-negative) predictor (Miconi 2017).

[8] This, of course, does not constitute a knock-down argument against the technical view of algorithmic fairness. Yet, as I have argued that it is less open to the critical reflection on the ideas of 'fairness'. So, it is less likely to be sensitive to the contentious nature of the definitions of fairness too.

In addition to the impossibility theorem, others have pointed to the inherent trade-off between fairness and accuracy in algorithms (see, e.g. Corbett-Davies *et al.* 2017; Berk *et al*. 2017). 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 accuracy in the case of the COMPAS recidivism algorithm in terms of fairness (in terms of disparate impact) and public safety, where optimizing for fairness measures is translated as the failure to detain the medium- to high-risk defendants who are more likely to commit violent crimes, and thereby threatening the public safety.

For those who value public safety, fairness measures that significantly reduce public safety thus will not be acceptable. Moreover, they might argue that fairness measures cannot be *genuinely* fair when these measures reduce *their* public safety, as optimizing for fairness impose risks—or, more precisely, the risks of harm—on people for the benefits of the defendants.[9] In other words, prioritizing fairness could unfairly put some members of the public at the risk of harm from violent crimes.[10] 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 accuracy points to the fact that whether the choice of fairness measure will be deemed 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 societal values in the process.[11]

The impossibility theorem and the inherent trade-off between fairness and accuracy, therefore, raise the following questions: if researchers 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*?

In answering these questions, Narayanan (2018) helpfully reminds us that the different fairness measures can be understood as representing the interests of different stakeholders affected by an algorithm. For example, in the case of the COMPAS recidivism algorithm, judges

[9] There is an important distinction between *actualized* harm and *risk* of harm to be made in discussing fair distribution of risk, see Hayenhjelm (2012), Hayenhjelm & Wolff (2012). Yet, the debate on distributive justice and risk is out of the scope of current paper, but my argument here only relies on the assumption that the distribution of risk and potential benefits is in fact an issue of fairness.

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

[11] In this respect, the increasing number of researchers being more explicit about the values and/or normative grounds of various definitions of fairness is a welcoming trend in research on algorithmic fairness (e.g. Dwork *et al*. 2012; Friedler et al. 2016; Berk et al. 2017; Narayanan 2018).

and parole officers will focus on the (positive) predictive value of the algorithm, i.e. how many correct instances of recidivism can it successfully identify; but, they will also want to ensure irrelevant and/or sensitive features, such as the defendants' race, do not directly affect the prediction; whereas for the defendants, especially those in the protected (minority) group, their concerns are about the chance of being mistaken by the algorithm as medium- or high-risk, and thereby facing more severe penalty due to the algorithm's error, this group of individuals will demand the chance of being misclassified not to be significantly greater than of other groups (Narayanan 2018).

Making explicit the relation between stakeholders' interests and fairness measures is of tremendously importance because it invites us to go beyond seeing algorithmic fairness merely as a technical task of implementing some pre-given ideas of 'fairness' into algorithms. As the choice of *any* fairness measure will inevitably favor the interests of some groups of stakeholders over other groups of stakeholders, thereby potentially benefiting some while harming others. So construed, algorithmic fairness is not only about designing and implementing algorithms that satisfy some fairness measures, but about *which ideas of 'fairness'* and *what other values* should be considered and accommodated within an algorithm. This, in turn, poses a significant ethical and political challenge to those who decide which fairness measures and what other values an algorithm is to install.

Moral philosophers have long argued that imposing significant risk on people without their consent is *prima facie* wrong, and that consent is morally necessary for risk imposition on individuals (MacLean 1982; Teuber 1990). When an algorithm is devised to make high-stake decisions *about* or *for* individuals, those who are affected by the algorithm can legitimately question whether the choice of specific fairness measure and the balance between fairness and performance put them at significant risk and insist that their consent is necessary in order for the choice to be morally defensible.[12]

Similarly, political philosophers and political scientists have argued the all-affected principle, as an ideal of democratic society, dictates that those who are significantly affected by a decision ought to be included in the decision-making either directly or indirectly (Dahl 1990, 49; cf. Whelan 1983). In a democratic society, people who are affected by the choice of fairness measure and of the balance between fairness and performance of an algorithm, therefore, ought to have a say in the decision-making. Yet, this is complicated by the fact that different fairness measures and balances may represent *conflicting* interests of different groups of stakeholders, and each of them will see different choices of fairness measure and balance as the 'right' one. To settle on an understanding of fairness and to strike a balance between fairness and performance of an algorithm, therefore, is not merely a technical task, but a *political* task that requires researchers to consider and accommodate diverse, conflicting interests of those who are affected by the algorithm, and it is a task that should be undertaken

[12] Hansson (2006) has forcibly questioned the applicability of (informed) consent in non-individualistic contexts. Here, the discussion is by no mean an argument for the role of consent in morally justifying imposition of risk from algorithms, it is merely an example of the type of ethical questions that could be raised.

by the researchers *with* the people. In short, the impossibility theorem and the inherent trade-off between fairness and accuracy call for an opening-up of the definitions of fairness for public discussion.

An Accountability for Reasonableness Framework for Algorithmic Fairness

By now, I hope I have made a strong case against the sufficiency of the technical view of algorithmic fairness. Setting aside the question about whether and which uses of algorithms in high-stake decision-making are morally permissible, if we think that algorithms can be used in some contexts, creating fair(er) algorithms should remain an imperative.[13] In the remainder of this paper, based on Daniels and Sabin's (1997, 2008) accountability for reasonableness (AFR) framework, I shall outline a framework for algorithmic fairness that accounts for the *political* nature of algorithmic fairness.

Two caveats must be mentioned before I elaborate the AFR-inspired framework for algorithmic fairness. First, I shall assume that there is intractable disagreement among various groups of stakeholders over the question of which conceptions of fairness, fairness measure, and balance between fairness and accuracy are the *right* one, because stakeholders have diverse, conflicting interests. If there is an uncontroversial understanding of fairness and of the priority between fairness and other societal values for the algorithm in question, then the technical view should be sufficient.[14] Second, I shall also assume that the interests expressed by various groups of stakeholders and their preferences for specific ideas of 'fairness', fairness measure, and balance between fairness and accuracy are morally and politically justifiable.[15] These two assumptions do not only reiterate the difficulty to view algorithmic fairness as technical issue, which requires some pre-agreed ideas of fairness and priority between fairness and other societal values, they also underscore a peculiar condition of liberal democratic society, that is—it is characterized by the *pervasiveness of reasonable disagreement*, or, as John Rawls (1993) calls it, *the fact of reasonable pluralism*.[16] Since reasonable disagreement is ineliminable in a liberal democratic society, we can only aim at *reducing disagreement* and *accommodating differences*.[17] It is against this background I

[13] If one considers *every* use of algorithmic decision-making to be morally impermissible, then concerns over fairness in algorithms will cease to exist. In other words, the project of creating fair algorithms presupposes some uses of algorithms to be morally permissible.

[14] Yet, even if there is *no* disagreement among different groups of stakeholders, I take it that the approach I outline can *enhance* the 'fairness' of the choice.

[15] My discussion *only* requires there are at least *some* choices that are equally justifiable, and thereby leading to the problem for justifying one justifiable choice over another *equally* justifiable choice. Hence, this is *not* to claim that *any* interests or the choices based on those interests can be justified.

[16] 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, xvi).

[17] Rawls argues that despite the 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 conceptions of justice with their own reasons and from their own moral point of view (cf. Rawls 1993, 134). Yet, the agreement on political conception of justice is necessarily thin, and thus it

introduce Daniels and Sabin's AFR framework to the context 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 decisions over appropriate ideas of 'fairness', fairness measure, and balance between fairness and accuracy in designing and implementing fair algorithms.

Daniels and Sabin's AFR is developed to respond to the problem of limit-setting in healthcare contexts.[18] They argue that heathcare is a fundamental human 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, job opportunities, etc. Hence, 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 others (Daniels & Sabin 2008, 13-24). Moreover, they also argue that there is neither consensus amongst people on the how the limits 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 others in a democratic society (Daniels & Sabin 2008, 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 "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 & Sabin 2008, 4), because there is no pre-agreed or universally accepted normative standard can be invoked to justify the limit on healthcare goods (or other human goods). For Daniels and Sabin, the *normative* question of limit-setting thus has to be reformulated into the 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 & Sabin 2008, 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 & Sabin 2008, 26). The shift towards procedural justice, namely to identify the conditions where decisions are morally and politically acceptable on the grounds of legitimacy and reasonableness of the decisions, allows us to proceed with limit-setting in absence of a consensus for normative standard. It also highlights Daniels and Sabin's commitment to the democratic ideal of the all-affected principle.

is insufficient to supply fine-grained normative principle to settle substantive issues, e.g. prioritizing the interests of different groups of stakeholders (cf. Daniels 1993).

[18] Daniels and Sabin first proposed AFR in Daniels & 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 therefore I shall not attempt to survey the extensive discussion on AFR. My discussion of AFR refers primarily to Daniels & Sabin (2008), which incorporates the earlier works of AFR and presents the most systematic account of it. However, I shall also refer to the earlier works on AFR when I deem them more focus on a specific point under discussion.

It is useful to elaborate the parallels between the problem of limit-setting in healthcare contexts and the problem of algorithmic fairness, as their similarities help to further showcase why AFR is fitting in the context of algorithmic fairness. First, both the problem of limit-setting and the problem of algorithmic fairness are based on the existence of reasonable disagreement in liberal democratic societies. AFR eschews the search for a pre-agreed or objective 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 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. Yet, since decisions about healthcare goods—and, in the case of algorithmic fairness, decisions about fairness measures—have to be made, AFR's response is to spell out the conditions to ensure the decisions to be acceptable to those who are affected by them through inclusion and accommodation of their views and voices. Finally, for Daniels and Sabin (2000; also, see 2008, 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 private organizations presumably 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, makers of high-stake decisions—either directly by designing and implementing an algorithm, or indirectly by proposing specific ideas of 'fairness', fairness measure, and related optimization techniques for the algorithm. Here, pre-determining or presuming the meaning of fairness by researchers will be morally problematic, as doing so risks neglecting the views and voices of those who are affected by the algorithm. AFR overcomes this risk by specifying the conditions for decision-making where people's views and voices will be accounted for.

It is these similarities between the problem of limit-setting and the problem of algorithmic fairness as well as AFR's potential to address the peculiar background condition of liberal democratic societies that make it a suitable framework for algorithmic fairness. We may even view the choice of fairness measure and the balance between fairness and accuracy 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 liberal democratic societies.

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 are originally intended for healthcare contexts, I have amended the four conditions to make them directly applicable to the case of algorithmic fairness: [19]

[19] The formulation of the four conditions I quoted is slightly different from the one presented in Daniels & Sabin (2008, 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.

> 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 [algorithmic fairness] or the best way to meet [claims] of the defined population under reasonable resource constraints. Specially, 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, 144-145; original emphasis).

Daniels and Sabin argue that the Publicity condition in AFR ensures transparency of decisions and decision-making processes, and it allows people to observe whether the decision-makers are coherent and consistent in their decision-making (Daniels & Sabin 2008, 12, 46-47). Making the reasons for decisions public can also force the decision-makers to clarify their rationales and relate them to the people, as the reasons are open to scrutiny and debate by the people, which, in turn, can contribute to the quality of public deliberation and facilitate social learning (Daniels & Sabin 2008, 47-49). In doing so, the decision-making processes that meet the Publicity condition demonstrate that the decision-makers are *principled* and *responsive* to the people, in particular to those who are affected by their decisions, thereby grounding legitimacy to the decision-makers.

In the case of algorithmic fairness, the Publicity condition requires publicizing both the choice of fairness measure and the rationales for adopting such a choice; and, this is indeed what Northpointe has done after being criticized by ProPublica. Although the *right* fairness measure to be used in criminal risk assessment algorithms remains undetermined, it is correct to assert that publicizing the current fairness measure in the COMPAS recidivism algorithm and the rationales for it does contribute to social understanding of the problem of algorithmic fairness in criminal risk assessment algorithms, and the same should hold for other algorithms too. More importantly, as I noted earlier, algorithmic bias is difficult to detect, so it will be difficult for individuals to know *how* an algorithm will affect people and *who* it will affect. If the purpose of the Publicity condition is to enhance public deliberation and social learning, I shall add that the consequences of an algorithm and to which group the algorithm will affect ought to be made plain to the public in *non-technical language*, especially because different fairness measures have varying implications to different groups (Chouldechova & 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 from the use of algorithms.[20] Here, my addition to the Publicity condition entails that if the industry fails to explain in layman's terms the consequences and distributional implications of an algorithm, or of the choice of specific fairness measures, the use of algorithms or the choice of fairness measures will be considered to be 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 terms of cooperation" (Daniels 2010, 145), but it has been subjected to a number of criticisms (see, e.g. Friedman 2008; Lauridesn & Kippert-Rasmussen 2009; Ford 2015; Badano 2018). For example, the important notion of "fair mind" people is left undefined 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 & Sabin 2008, 44-45); and, 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.

In his 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, 18).

Badano borrows insights from Nagel (1979, 1991) and Scanlon (1982) to argue 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, that is—to strive 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 people 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, 11-14).

Of course, the Full Acceptability condition in itself does not present a standard to adjudicate between competing claims, but it 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 (most) vulnerable, and objects the use of impersonal reasons, e.g. overall efficiency of the society, to override their claims. In the

[20] Veale and Binns (2017) rightly point out that there are practical difficulties for private organizations in explicating 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. They have provided three responses to this problem for private organization. While I cannot examine their proposals in detail, I think the proposed responses are compatible with the AFR-inspired framework I develop in this paper.

case of algorithmic fairness, the Full Acceptability condition then requires examining more closely the claims of those who are, or will be, negatively affected by the use of algorithms in high-stake decisions, particularly the most vulnerable (see, e.g. Woodruff *et al*. 2018), and deciding on fairness measure on the basis that the choice will be most acceptable to people to whom the choice is least acceptable.[21]

As the 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 on-going 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. In other words, the Revision and Appeal condition is necessary to cope with the rapidly changing social and technological environment. In effect, without a proper means to review and revise previous decisions, the cost of mistakes will be excessively high and therefore hinder decision-making. In addition, good mechanisms for challenge and dispute resolution can strengthen the legitimacy of decision-making and contribute to social learning of the problem at hand, for 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 & Sabin 2008, 58-59). The need to review and revise decisions in the case of algorithmic fairness is even more pressing, as the use of algorithms in high-stake decision-making is still in its early days, we can readily expect research findings that disrupt our preconception about the choice of fairness measures and its consequences, or novel techniques to automate decision-making.

Finally, the Regulative condition is proposed to ensure private organizations' adoption of the Publicity condition, the Relevance condition (or, as proposed by Badano, the Full Acceptability condition), and the Revision and Appeal condition. In the context of algorithmic fairness, the Regulative condition entails the necessity of regulations and public agencies to enforce the three conditions on private organizations.

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 to see the answer to be emerged from public deliberation. Here, it is the contestable nature of the idea of 'fairness' in the problem of

[21] 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 therefore 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 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 suppression of differences (Young 2000, 221-228).

algorithmic fairness and the opportunity to overcome the view that algorithmic fairness is primarily technical that make AFR particularly suitable to approach algorithmic fairness. The AFR-inspired framework for algorithmic fairness aims to open up the question of 'fairness' in algorithmic fairness to the public, especially to those who are affected by algorithms, and it attempts to ground the choice of the definition of fairness for the designing and implementing fair algorithms with democratic communication. In this respect, the normative foundation of the AFR-inspired framework rests ultimately on deliberative democracy (Daniels & Sabin 2008, 34-36; also, see Gutman and Thompson 1996, 2004; Habermas 1996; Young 2000).[22]

It is thus worth to reemphasize that the AFR-inspired framework does not just offer a different *substantive* idea of fairness for the technical view of algorithmic fairness, 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 the algorithmic decision-making: it requires researchers and the people to be public about their decisions and rationales for their decisions, 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 AFR's insistence on decision-making should be viewed as an on-going process that can be defeated when new knowledge and technologies come into the picture, that is—decision-making should be open to *new* reasons.

Here, the emphasis on the exchange of reasons and the view that decision-making should be regarded as an on-going process can be helpfully illustrated by a contrast with Grgić-Hlača *et al.*'s (2018) recent proposal for a *technical* approach to *procedural justice* for algorithmic fairness. They crowdsource individuals to vote on the features they consider to be fair to include in algorithmic decision-making and design 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 AFR-inspired framework, because it is based on an aggregation of preference, which does not involve genuine *exchange* of reasons. For the AFR-inspired framework, it is the *reason-giving* and *reason-responding* in the exchange of reasons that demonstrate the respect to individuals' views and voices and the recognition of their differences. Moreover, it is through the exchange of reasons that

[22] The more fundamental questions for the AFR-inspired 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-inspired 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, etc. 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 the AFR-inspired framework. For a review of the prospect of deliberative democracy, see Curato *et al.* (2017)

different parties involved learn more deeply about the problem of algorithmic fairness and learn from those who are adversely affected by the uses of algorithms. From this point of view, mere aggregation of preference will be insufficient.

Concluding Remarks

In this paper, I attempt to show that the technical view of algorithmic fairness is insufficient, because of the contentious nature of the ideas of 'fairness' and the fact that the decisions over fairness measure and the balance between fairness and accuracy are about *competing* values. One of the main contributions of this paper, therefore, is the explicit reconceptualization of algorithmic fairness essentially as a *political* issue.[23] Note that my claim is not merely there is a political dimension for algorithmic fairness, but a more radical claim that the problem of algorithmic fairness is first and foremost *political* and the technical task only comes in as secondary. Hence, the problem of algorithmic fairness needs to be resolved by political means, and I have proposed a version of AFR to this task.

The AFR-inspired framework I proposed in this paper requires the industry and research community who develop fair algorithms to account for the interests of those who are affected by the algorithms through making public the ideas of 'fairness', fairness measure, and balance between fairness and other societal values in designing and implementing the algorithms, grounding their decisions with reasons that are acceptable by the most adversely affected, and being open to adjustments in light of new reasons. These requirements have been echoed in various sets of ethical and governance principles for algorithm design and implementation (see, e.g. Diakopoulos *et al.* n.d.; USACM 2017; Reisman et al. 2018).[24] This should not be surprising, as the four conditions in AFR succinctly capture the basis of what is it for decision-makers to be accountable and their decisions to be legitimate. More importantly, the AFR-inspired framework offers *philosophical* and *normative* grounds for these requirements, and thereby supplementing the existing ethical and governance principles.

Yet, my discussion of the AFR-inspired framework is far from complete, nor have I provided a full defense of the framework in the context of algorithmic fairness. Particularly, the theoretical and practical questions related to the Full Acceptability condition (or, the

[23] Binns (2017) 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.

[24] For example, the four policy goals of Algorithmic Impact Assessments (AIA), one of the most elaborated frameworks proposed by AI Now Institute, are: "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 […]; 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, 5).

Relevance condition) require further exploration, and the methods of public deliberation and their limitations that could truly enable agreement and accommodate differences also require substantial future works. [25] The goal of my discussion of the AFR-inspired framework, therefore, is a modest one, that is—by foregrounding the similarities between the problem of limit-setting in healthcare contexts and the problem of algorithmic fairness, I offer a reformulation of the problem of algorithmic fairness as a problem of limit-setting, and demonstrate that AFR is a promising framework for the problem. There are a number of unfinished business for a fully elaborated AFR-inspired framework. Yet, I consider this paper to have made explicit the problem of algorithmic fairness first and foremost as a political problem and provided a plausible framework to account for it, thereby inviting future works to be done in this direction.

Reference

ACM US Public Policy Council [USACM] (2017). Statement on Algorithmic Transparency and Accountability. Available Online at: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

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

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

Arneson, R. (2018) Four Conceptions of Equal Opportunity. *The Economic Journal*, Early View. Available Online at: 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

Berk, R. Heidari, H. Jabbari, S. Kearns, M. Roth, A. (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv preprint*, arXiv:1703.09207

Binns, R. (2018) Fairness in Machine Learning: Lessons from Political Philosophy. *Journal of Machine Learning Research* 81, 1-11

Binns, R. (2017) Algorithmic Accountability and Public Reason. *Philosophy & Technology*, Online First. Available Online at: https://doi.org/10.1007/s13347-017-0263-5

boyd, d. Crawford, K. (2012) Critical Questions for Big Data. *Information, Communication & 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 & Society* Jan-Jun 2016, 1-12.

[25] For example, see Ney and Verweij (2015) for an excellent discussion of different methods to engage the public, and to accommodate the normative principle 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.

Chouldechova, A. (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2), 153-163.

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)*, Halifax.

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. *arXiv preprint*, arXiv:1701.08230

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. (2012) Reasonable Disagreement about Identified vs. Statistical Victims. *Hastings Center Report* 42 (1), 35-45.

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 Edition. 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. Available Online at: http://www.fatml.org/resources/principles-for-accountable-algorithms

Dieterich, W. Mendoza, C. Brennan, T. (2016) COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpoint Inc. Available Online at: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

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* (pp. 214-226). Cambridge, Massachusetts.

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. (2018) A Comparative Study of Fairness-enhancing Interventions in Machine Learning. *arXiv preprint*, arXiv:1802.04422

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. In Proceeding of The Thirty-Second AAAI Conference on Artificial Intelligence (pp. 51-60), New Orleans.

Gutmann, A. Thompson, D. (1996) *Democracy and Disagreement*. Cambridge, MA: 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.

Hansson, S. O. (2006) Informed Consent Out of Context. *Journal of Business Ethics* 63 (2), 149-154.

Hagendijk, R. Irwin, A. (2006). Public Deliberation and Governance: Engaging with Science and Technology in Contemporary Europe. *Minerva* 44 (2), 167-184

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. (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint*, arXiv:1609.05807

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. (2017) Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, Online First. Available Online at: https://doi.org/10.1007/s13347-017-0279-x

Lipton, Z. C., Chouldechova, A. McAuley, J. (2018) Does Mitigating ML's Impact Disparity Require Treatment Disparity? *arXiv preprint*, arXiv:1711.07076

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.

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. Available Online at: https://speak-statistics-to-power.github.io/fairness/old.html

Mittelstadt B.D. Allo, P. Taddeo, M. Wachter, S. Floridi, L. (2016) The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* Jul-Dec 2016, 1–21.

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

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

Narayanan, A. (2018, February) 21 Fairness Definitions and Their Politics. Tutorial presented at: *The Conference on Fairness, Accountability, and Transparency (FAT*)*, NYC. Available Online at: https://www.youtube.com/watch?v=jIXIuYdnyyk

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.

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

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

Reisman, D. Schultz, J. Crawford, K. Whittaker, M. (2018) *Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability*. Available Online at: https://ainowinstitute.org/aiareport2018.pdf

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, *arXiv preprint,* arXiv:1706.09976

Temkin, L. (2017) The Many Faces of Equal Opportunity. *Theory and Research in Education* 14 (3), 255-276.

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

Veale, M. Binns, R. (2017) Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. *Big Data & Society* Jul-Dec 2017, 1-17.

Wallach, W. Colin, A. (2009) *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

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). Montreal.

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.