Delft University of Technology

MinMax fairness
from Rawlsian Theory of Justice to solution for algorithmic bias

Barsotti, Flavia; Koçer, Rüya Gökhan

DOI
10.1007/s00146-022-01577-x

Publication date
2022

Document Version
Final published version

Published in
AI and Society

Citation (APA)
Barsotti, F., & Koçer, R. G. (2022). MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias. *AI and Society*, *39 (2024)*(3), 961-974. https://doi.org/10.1007/s00146-022-01577-x

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

AI & SOCIETY (2024) 39:961–974
https://doi.org/10.1007/s00146-022-01577-x

ORIGINAL PAPER

MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias

Flavia Barsotti[1,2,3] · Rüya Gökhan Koçer[1]

Received: 8 April 2022 / Accepted: 26 September 2022 / Published online: 30 November 2022
© The Author(s) 2022

Abstract
This paper presents an intuitive explanation about why and how Rawlsian *Theory of Justice* (Rawls in A theory of justice, Harvard University Press, Harvard, 1971) provides the foundations to a solution for algorithmic bias. The contribution of the paper is to discuss and show why Rawlsian ideas in their original form (e.g. the *veil of ignorance*, *original position*, and allowing inequalities that serve the worst-off) are relevant to operationalize fairness for algorithmic decision making. The paper also explains how this leads to a specific *MinMax fairness* solution, which addresses the basic challenges of algorithmic justice. We combine substantive elements of Rawlsian perspective with an intuitive explanation in order to provide accessible and practical insights. The goal is to propose and motivate why and how the *MinMax fairness* solution derived from Rawlsian principles overcomes some of the current challenges for algorithmic bias and highlight the benefits provided when compared to other approaches. The paper presents and discusses the solution by building a bridge between the qualitative theoretical aspects and the quantitative technical approach.

Keywords Algorithmic bias · Fairness · Rawlsian Justice · Ethics · AI systems

1 Introduction

The last decade has highlighted an accelerating use of Machine Learning (ML) and Artificial Intelligence (AI) in a wide range of domains, especially in finance, justice and security as a tool to facilitate or even replace human-based decision-making: this has revealed the intimate links between politics, ethics, and computer science. Looking at justice in society, the realization that AI may actually generate or reinforce inequalities due to its potential to discriminate individuals on the basis of sensitive characteristics (such as ethnicity, race and gender) triggered the demand for ethical use of AI (European Commission 2019; EBA 2020). Consequently, besides a multitude of guidelines and codes of conduct, the European Commission has released key supranational regulations related to the use of data (European Commission 2018) and AI systems, e.g. the draft AI Act (European Commission 2021). We might expect that in the near future the use of AI will be strictly regulated in high-risky domains (for example, credit-provisioning and security). Therefore, for an industrial user of AI systems operating in such domains it is now necessary to define justice and fairness in tangible ways in order to design AI systems and applications accordingly. Thus, corporate actors need to answer questions of ethical and ultimately philosophical nature in order to leverage on the advancement of technology.

With a focus on fairness and algorithmic decision making, the typical first step in implementing a fairness solution is to select a specific fairness definition. This initial decision is the most difficult one both from technical and ethical points of view: there are many fairness definitions (Dwork et al. 2012; Hardt et al. 2016; Joseph et al. 2016) that can be operationalized through various technical approaches; but any one of these definitions can easily be contested with convincing arguments. Moreover, fairness definitions do not help people that are vulnerable on

✉ Flavia Barsotti
flavia.barsotti@ing.com; f.barsotti@uva.nl; f.barsotti@tudelft.nl

Rüya Gökhan Koçer
ruya.kocer@ing.com

[1] ING Analytics, Amsterdam, The Netherlands

[2] IAS (Institute for Advanced Study), University of Amsterdam, Amsterdam, The Netherlands

[3] DIAM (Delft Institute of Applied Mathematics), TU Delft, Delft, The Netherlands

the basis of attributes that might be not accessible or not yet known: this is for example the case for disabilities or peculiar mental characteristics. Therefore, any given fairness definition would be arbitrary not only because it can be objected by context-dependent substantive arguments, but also due to its inability to address all potential instances of unfairness. This issue and its impacts become even more relevant when these attributes are considered commonly on the basis of "group identities" derived from pre-defined 'known' "protected" characteristics, such as race or ethnicity. Thus, despite the technical sophistication of different engineering solutions, the fundamental problem in algorithmic justice remains unanswered: any given fairness definition can be seen as arbitrary or based solely on convenience rather than any fundamental concern or argument about justice. As a consequence, the field of computer science seems to have reached an impasse similar to the one emerged in political philosophy in the mid-20th century: there are many definitions of fairness (e.g. justice) without any compelling reason for choosing any one of them in any given context. It was in response to this impasse that John Rawls had produced his seminal study *Theory of Justice* (Rawls 1971, 1999) and introduced the idea of *Justice as Fairness* (Rawls 1985, 2001).

Rawls was addressing the basic problem related to the fact that any notion of justice can be defended and objected by equally convincing arguments; thus, it was necessary to figure out a way to attain a universal definition that (1) would be acceptable by all people regardless of their convictions about justice, and (2) that would help all people regardless of their individual characteristics. Rawlsian contribution represents a key milestone for political philosophy. From the perspective of consistency and systematic premise, it provided a basis for what Rawls calls "intuitionist" theories of justice that had been invoked against various forms of utilitarianism. Not surprisingly, literature contains many sharp criticisms of Rawlsian theory, especially against his idea of a single conceptualization of justice. However, even for those who oppose his arguments, Rawls provides a clear argument against which others can position themselves (Freeman 2009).

Interestingly, the principles developed by Rawls in order find a unique solution for justice are quite relevant for algorithmic fairness nowadays. Indeed, recent literature on "fairness without demographics" approach has already recognized this connection (Lahoti et al. 2020; Hashimoto et al. 2018; Martinez et al. 1970). However, these studies explore different technical solutions for fairness while a motivation of the principles linked to Rawlsian ideas is not discussed in detail, leaving open the question on how exactly their solutions are derived from the principles of Rawlsian *Theory of Justice* (Rawls 1971). Consequently, these contributions do not directly address the problem of explaining why their solutions are fair.

In this context, the goal of this paper is to discuss the main principles introduced by John Rawls and show why they provide a solution for the current challenges of algorithmic bias in the form of *MinMax fairness* and link them to a specific formal definition. To achieve this, we combine the substantive discussion of Rawlsian principles with an intuitive formalization in order to provide accessible insights. Accordingly, we propose a definition of fairness on the basis of Rawls *maxmin principle*, aiming to maximize the minimum utility in its original form, and name it *MinMax fairness*, with the goal of minimizing the maximum errors made by the model. Our definition reads as follows: *'A model is fair if it does not make more systematic errors for any sub-group in the dataset compared to the others'*; in this way, we translate *MinMax fairness* into a concrete criteria to assess fairness in models, by focusing on the errors made by the model (see Sect. 4, Definition 4.1) and introducing a weighting idea to 'force' the model to correct them. The paper also explains why and how the Rawlsian solution overcomes some of the key challenges for algorithmic bias, highlighting the benefits it provides when compared to other fairness approaches (i.e. parity-based) in terms of technical engineering choices and ethical implications. In general, *MinMax fairness* is applicable in different contexts (e.g. health, banking, welfare, etc.), and does not have specific restrictions in terms of modeling approaches or application domain. In this paper we discuss the operationalization of *MinMax fairness* solution as fairness intervention in the particular context of credit decision modeling as illustrative case.

The paper is structured as follows: Section 2 introduces Rawlsian *Theory of Justice* principles and discusses them from both a societal and economic perspective by highlighting the link between the original principles and their potential usefulness for algorithmic fairness. Section 3 hints how the concrete implementation could be valuable to solve some open questions on algorithmic fairness. Section 4 provides an intuitive explanation of the mathematical solution linked to *MinMax fairness* and its operationalization via an illustrative example. Section 5 concludes.

2 Rawlsian *Theory of Justice*

This section introduces and discusses the main ideas behind Rawlsian *Theory of Justice* (Rawls 1971) and the concept of *Justice as Fairness* (Rawls 1971, 1985, 1999, 2001) relevant to build the proposed *MinMax fairness* solution for algorithmic bias. We first introduce a conceptualization of the main ideas related to *democratic equality* and *equal opportunity* principles and then discuss the concepts of *original position* for citizens and *veil of ignorance*.

2.1 *Democratic equality* and *equal opportunity*

One of the basic premises of Rawls' definition of *justice* is *democratic equality*. In his seminal work, Rawls states that people are all equal and deserve to lead their lives as free and equal citizens based on what they consider being a 'good life': in this respect, this goal should not be affected by arbitrary factors. The main Rawlsian argument is linked to the concept of *equal opportunity*: inequalities observed in society could be considered as "just" as far as people attain goals that differentiate them from other people on the basis of their own effort and personal free choices (Fleurbaey 1995; Barry 2017). However, according to Rawls, the concept of *equal opportunity* is a complex issue in society and suffers from a serious setback. Rawls observes that people do not commence their lives from identical circumstances. Many factors, often beyond their control, could create serious implications for the opportunities that they can encounter or hope to encounter during their lives. Sensitive characteristics such as race, ethnicity and gender are examples of factors that can deeply impact what people can achieve through their own efforts. Rawls considers the life circumstances that emerge due to advantages and disadvantages generated by such socially constructed categories as 'unfair'. And Rawls goes beyond this, hinting that socially constructed categories are not the only factor generating unfairness. According to Rawls, also natural talents may lead to unfair outcomes as these talents are also arbitrarily distributed: often, people cannot really choose to have a tendency or talent for a particular trait, such as high intelligence. These traits too are often distributed rather randomly by nature (Kymlicka 2002). These considerations related to potential "unfairness" affect the way in which Rawls rectifies the equal opportunity argument. Rawls argues that inequalities could still be permitted but this should happen in such a way that they should be beneficial for those who are mostly disadvantaged in terms of the arbitrary factors that affect people's opportunities. This approach is expressed by the *maxmin principle*, that could be explained as the idea of *'maximizing the gains (e.g. utility) for those people who are in the minimum, that is, "worst-off" position'*. It is important to notice that this conceptualization does not imply a simple scheme of allocating more resources for those with a disadvantage, but rather requires to think of best possible ways of organizing the resources in order to ensure the resulting arrangement would help these people (Daniels 2003). For instance, offering higher remuneration to people who are skilled enough to be a surgeon would help less fortunate people to be more healthy if the access to health care is assured. Surgeons with a higher remuneration would probably be able to perform their duties better with more time to spend for improving themselves in their trade and resting when necessary reducing potential negative consequences deriving from restless habits and low remuneration. In this example, the resource allocation would ensure that people in the "worst-off" position—in terms of either talent or health conditions—would benefit from the outcome. This is the essence of Rawlsian conceptualization of justice for the society as group of free and equal citizens.

2.2 *Original position* and *veil of ignorance*

What makes Rawls' principles of justice particularly appealing for algorithmic fairness is the way in which he justifies his approach. He states this by means of two instruments: the *original position* and the *veil of ignorance*. In this context, Rawls' principles could be considered to create a direct connection with some of the basic challenges related to algorithmic fairness, both from a theoretical and practical perspective and build an elegant solution. It is useful to briefly examine these two concepts and discuss how they help to link Rawlsian notion of justice to a solution for algorithmic fairness.

Rawls' *original position* is based on a counterfactual environment in which people are imagined to be in a condition where there is no political authority to ensure social existence. The purpose of this device is to detach people from the current organization of society and help them to think of ideal conditions that they would have preferred if they had a chance to organize political authority from scratch (Dworkin 1973; Clark 1993). The idea is to discover the basic principles that would be agreed upon independently by each individual, rather than figuring out a specific account for the entire political system. This counterfactual argument is built on the underlying assumption that *agents are rational* at the *original position*. Rawls argues that the *original position* would not be sufficient to derive principles of fairness and justice as people might take their own material, mental or physical conditions as they determine the principles of justice that need to be adhered to. It stands to reason to imagine that at least some people would endorse those principles that would generate a favorable position for them in accordance with their specific conditions. Obviously, the other people cannot be expected to accept such principles due to potential disadvantages that this would ensue for themselves. The concept of *veil of ignorance* allows Rawls to build a solution for this conundrum. This refers to an additional condition added to the *original position* ensuring that individuals are completely "blind" about their own material, mental or physical conditions in the political order (and thus society) that would emerge after they determine the principles of justice. According to Rawls, only after a *veil of ignorance* is put in front of people, the question of what should be the principles of justice that need to be used to organize a society makes sense (Roemer 2002). Rawls argues that, in a society with rational agents, once a *veil of ignorance* is combined with the *original position*, individuals would opt for the *maxmin principle*. In practice, people would prefer the available resources to be distributed in such a way that the resulting inequalities would be most favorable for those people that are disadvantaged due to arbitrary factors. The intuition can be explained as follows: for each and every individual the possibility of suffering from at least some disadvantages would be identical, at least from their vantage point blocked by the *veil of ignorance*; consequently, people would like to avoid creating a society in which they could potentially be in a disadvantaged position. In a way, the *veil of ignorance* has the function of converting all other people into a possible position in the society for a given individual ensuring that s/he will take everybody into account (Kymlicka 2002).

3 From Rawlsian *Theory of Justice* to algorithmic bias

An important open challenge for algorithmic decision making is preventing harm to group of individuals on the basis of sensitive characteristics. The current approaches used by practitioners and the scientific contributions are mainly focusing on finding the proper fairness definition and the best engineering choice to mitigate bias. In our view, the field of computer science seems to be in an impasse similar to the one in political philosophy in the mid-20th century. As discussed in the introductory section, there are many definitions of fairness (e.g. justice) without any compelling reason for choosing any one of them in any given context. In response to the impasse in political philosophy, Rawls produced his seminal study *Theory of Justice* (Rawls 1971, 1999) that we consider the fundamental basis for a *MinMax fairness* solution to algorithmic bias. Starting from the conceptualization of Rawlsian principle of *justice*, this section explains how to translate this conceptualization into a solution for algorithmic bias. This is done both at theoretical level and from an implementation point of view via an illustrative example. The discussion also highlights the benefits of this solution when compared to parity-based approaches.

3.1 Algorithmic decisions

Let us consider the case in which an algorithmic decision making process produces a specific outcome affecting a decision about individuals in the context of a credit risk application: as example, the algorithmic decision could be 'whether to grant or not a loan to a single individual'. Moreover, the algorithmic decision could be based on the outcome of a simple modelling framework or a more advanced analytic approach based on specific ML or AI techniques. Let us consider a case with advanced analytic techniques to illustrate the problem.

The information about the information about the past behavior of a large group of individuals is input for the predictive model and is stored in two distinct formats:

- *Features values* related to single individuals. These represent particular characteristics such as age, income, education etc., each recorded in a quantitative or quantifiable format. For each single individual, we could consider the collection of these specific features values as represented by a single data point in a multidimensional space (input value).
- *Target value observations* coming from the past. In this special case, the "default history", e.g. whether or not the person identified by a series of features has paid back the loan or defaulted.

Based on this set of information, an AI algorithm discovers the patterns in the multidimensional space generated by the *features values* in the form of a probability score, namely $s \in [0, 1]$, indicating which combinations of features are likely to lead to higher default probability. In this illustrative example, once all patterns are captured in the form of *model parameters*, then the algorithm would be ready to be used as a decision making device for any new loan application.

Let us now consider a loan application coming from an individual who never applied before and see how this would be treated. In this case, it is the first time the algorithm 'sees' this data point: based on the information contained in her/his features values, the algorithm would generate a probability score $\hat{s} \in [0, 1]$ for the applicant (e.g. a value linked to her/his default probability) by using the *resemblance* between the new input values and the established patterns already learnt (i.e. learnt from the data about the past behavior of a large number of individuals and stored as parameters values). From an ethical perspective, the entry point of *justice* and *fairness* into this picture is the question of which features (e.g. individual characteristics) should be used to determine the default probability of each single individual.

3.2 Algorithmic bias

It is widely acknowledged that multiple sensitive characteristics that are beyond the control of single individuals (e.g. gender, ethnicity, disability, etc.) should not be used as features and rather be 'protected'; in some cases this directly relates to specific legal requirements. This aims to ensure that algorithms do not take the patterns linked to these sensitive characteristics into account and generate harm towards individuals on the basis of this information. This practice is also referred to as '*Fairness through unawareness*' (see Sect. 3.3), because the goal is to consider attaining fairness by being 'unaware' of the specific sensitive characteristics.

However it is possible that the information about the protected attribute is conveyed partially or fully by some other features that are included; and thus the protected attribute information may still be used without being noticed. As a consequence, bias may be rendered invisible rather than being mitigated.

Empirical evidence shows that algorithms are capable of learning and using the information about such sensitive characteristics even when they are not explicitly used. This might happen since the algorithm can implicitly 'see' such attributes in particular combinations of other features due to proxies and correlation effects (Pedreschi et al. 2008; Kusner 2017; Srivastava et al. 2019).

Let us imagine that the information associated with such characteristics is implicitly discovered by the algorithm: at this point, the usual logic underlying the outcome of the decision would be as follows. If the majority of those individuals who defaulted in the past do have a particular sensitive characteristic value (e.g. gender), then the majority of those having the same particular sensitive characteristic value would be classified as potentially defaulting on the loan. Consequently, any combination of other features that betrays this sensitive characteristic value would be used to *generate a bias* against those people having a similar set of sensitive characteristics but would not have defaulted. A similar logic would also apply in a positive sense: if the majority of those who did not default has a particular sensitive characteristic value (e.g. gender), then the algorithm would associate the majority of those who have this value to 'no default'; and in the process would grant loans to those people who would actually default. In both cases, a *false generalisation* would be made by the model with adverse consequences for the individuals impacted by this.

Figure 1 illustrates this logic by showing how potential harm can be generated for a given set of data points. Here imagine we look at the training dataset of a model that evaluates credit applications. The dashed sets indicate the model's predictions referring to the data points belonging to them. These data points are associated, respectively, to: 'prediction of default' or 'prediction of no default'. The filled/empty shape shows what had been in reality the outcome of the loan repayment process for a single data point: an empty shape refers to 'default', and a filled shape refers to 'no default', e.g. the person has fully repaid the loan. The majority of data points represented as triangles have not defaulted, while the majority of data points represented as circles have defaulted. However, there are three triangles—despite being triangles—who have defaulted, and there are three circles—despite being circles- who have repaid. Looking at the predictions, the model ignores these intricacies and wrongly predicts that three defaulting triangles would actually repay; and again wrongly predicts that three repaying circles would default. This happens because algorithms try to generalize on the basis of the patterns observed for the majority of data points in the training set, and attribute an outcome to a data point on the basis of the outcomes of the data points that mostly resemble to it. This is, in its essence, the source of errors and at the same time the reason for bias. A discrimination problem arises here: on the basis of being triangles, the three repaying triangles are discriminated negatively; and the same happens to the three defaulted circles, which in turn suffer from positive discrimination (they will suffer because they will not be able to pay back the credit they will get with adverse consequences). In this example these six data points that suffer from bias would be associated to systematic errors. In terms of magnitude, these errors will be the highest ones.

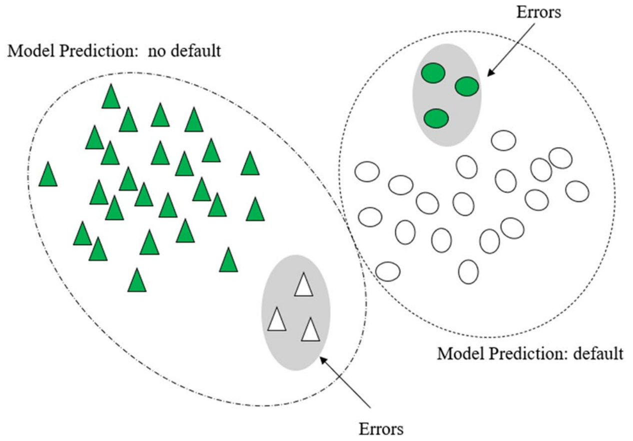

Fig. 1 Model predictions, default rates and errors. Data points in the training set. The shape of triangles or circles refers to a specific characteristic/feature; the filled/empty shape refers to a 'default' or 'no default' status. A filled shape refers to a data point associated with 'no default'; an empty shape refers to a data point associated with 'default'. Dashed sets identify the two alternative model's predictions: 'no default' and 'default'. Grey areas identify which data points are suffering from errors. In this illustrative case, empty triangles and filled circles are suffering from false generalisation resulting in specific errors made for these data points

The essence of the problem is that one individual should not be held accountable based on the actions observed in the past from those individuals who resemble him/her on the basis of characteristics that are arbitrary and sensitive. In other words, predicting one's future ability to repay a loan on the basis of what the majority of the others with the same sensitive characteristic have done in the past is tantamount to ignoring one's free will and individuality. This becomes even more relevant when considering that such characteristics are often sensitive and essentially 'unchosen' by each individual. In practice, this might directly lead algorithms to *generate de-facto discrimination towards specific groups* that is not desirable from an ethical perspective. Going back to the foundations defined by Rawls for *maxmin principle*, we show how a technical solution to algorithmic bias can be attained by means of similar arguments applied to the context of algorithmic decision making (e.g. bias/errors, sensitive attributes, fairness, etc.).

3.3 Attaining algorithmic fairness

We can classify the most common approaches to attain algorithmic fairness into two main streams: *Fairness through unawareness* and *Fairness through intervention*, described below.

1. *Fairness through unawareness*. As discussed in Sect. 3.2, this approach aims to attain algorithmic fairness by excluding the features that provide sensitive attribute information from being direct input for the model (thus being deliberately 'unaware'). However, in practice, it has already been established that this is not solving the bias problem (Dwork et al. 2012). Even if sensitive attributes are not directly entering into the model, it may happen that the information about the protected attributes is conveyed partly or entirely by some other features that are included. As a result, the protected attribute information may still be used without being noticed, and thus bias may be rendered invisible rather than being mitigated. For this reason, *Fairness through intervention* is considered a more appropriate direction.
2. *Fairness through intervention*. This approach aims to attain algorithmic fairness by directly introducing technical algorithmic interventions either during the model development phase (i.e. in-processing[1]) or afterwards as an additional layer (i.e. post-processing[2]). Within a model development pipeline, in order to detect and mitigate bias, alternative fairness interventions techniques are available and used in practice. Most commonly, to the best of our knowledge, fairness interventions are implemented by means of a *Parity-based approach*, aiming to achieve a form of parity of outcomes between different groups. Below we describe the main idea behind this intervention and present what we consider a novel direction for fairness intervention, namely *Accuracy-based approach*, to overcome the current challenges.

Parity-based approach consists of well-defined steps, e.g. assuming that the protected attribute information is available (e.g. ethnicity, gender), identifying privileged and unprivileged groups based on protected attribute information, selecting a particular fairness definition. The aim is to attain a parity between privileged and unprivileged groups on the basis of the chosen fairness definition (Verma and Rubin 2018; Bellamy et al. 2019; Haas 2020; Dwork et al. 2012), which is context dependent, and potentially arbitrary.

Rather than focusing on parity, we propose to go for a direction that we call *Accuracy-based approach*, whose main goals are described below.

Accuracy-based approach aims to increase the performance of the model for the protected groups on the basis of the potential of the model to make systematic errors for these groups. This focuses on improving the performance of the model in specific regions of the model space where there are clusters of data points that are most likely to represent vulnerable people with protected characteristics suffering from systematic errors. This is de-facto an operationalization of the following fairness intuition: *'A model is fair if it does not make more systematic errors for any sub-group in the dataset compared to the others'*. We refer to Sect. 4 (Definition 4.1) for a dedicated technical discussion of *MinMax fairness* definition and to Kim et al. (2019); Martinez et al. (1970) for examples from literature.

The most commonly used solution to the problem of preventing bias and discrimination considered by practitioners is ensuring a concept of *parity* among groups in accordance with a specific definition of fairness. This means that—on average—those groups that are defined by a sensitive characteristic do not receive privileged or unprivileged treatment from the algorithm on the basis of a particular definition of fairness. The on-going scientific debate is mainly focused around how to identify the proper fairness definition and how to address the fairness-performance trade-off. We can identify multiple definitions of fairness (Hardt et al. 2016; Joseph et al. 2016; Kearns et al. 2018), capturing a broad range of different legal, philosophical and social perspectives. As relevant issue, we might have cases in which, after a specific fairness intervention, a model is considered 'fair' on the basis of a given fairness definition: this is achieved at the cost of reduced model accuracy (Haas 2020; Dwork et al. 2012). In Aler Tubella et al. (2022), we show that the technical choice of a specific fairness intervention may have relevant implications in terms of which data points will be affected by the specific technical solution. The paper reveals how assessing the engineering choices in terms of their ethical consequences can contribute to the design of fair models and the related societal discussions. In this spirit, we

[1] In-processing refers to all methods that incorporate a fairness definition into the algorithm design and optimize it accordingly (as an example, see (Bellamy et al. 2019)).

[2] Post-processing methods refer to all techniques that intervene on the predictions produced by a model (without interfering with the algorithm) in order to attain unbiased outcomes (as an example, see (Bird et al. 2020)).

propose the following assessment of parity-based approach and a *MinMax fairness* solution for algorithmic bias based on Rawlsian conceptualization of justice, as example of accuracy-based approach. Section 4 provides an intuitive explanation of the technical and mathematical choices to build *MinMax fairness* solution and provides an illustrative example of implementation.

Parity-based approach to fairness suffers at least from four deficiencies:

(1.P) There are multiple ways to define fairness and no a-priory right fairness definition for each specific case. It is not easy to justify any fairness definition as the correct one for specific circumstances and always possible to object the chosen definition. Thus, the very first step of the process to build fair algorithms appears rather arbitrary.

(2.P) Parity-based definitions depend on the identification of groups on the basis of 'known' sensitive characteristics (e.g. race, gender, ethnicity). There are cases in which it is not possible to know—and often legally not permitted to know—some sensitive characteristics about individuals (Yang and Dobbie 2020; Andrus et al. 2021). As a consequence, often times parity-based fairness definitions are not implementable, even when having the most suitable one for the specific context. From a fairness implementation perspective, parity-based approaches have a clear limitation: they only allow, at best, to address existing bias appearing on the basis of 'known' sources.

(3.P) The implementation of parity-based fairness requires parity of outcomes between different groups; for the sake of equality, this intervention aims to reduce the difference between advantaged and disadvantaged groups, for example in terms of number of wrong predictions. Given how parity-based works, ensuring parity of outcomes often implies decreasing the utility of a model for a specific advantaged group, without necessarily increasing the utility for the disadvantaged groups. As a matter of fact, the information about the privileged group reaching a 'worst-off' position would not change the absolute material conditions of the unprivileged group. A parity-based approach, when implementable, enables to increase fairness at the cost of reducing accuracy.

(4.P) When parity-based is applicable and is used to ensure fairness on the basis of a specific 'known' attribute, it might happen that this generates (unintentionally) unfairness on the basis of other attributes, due to intersectionality issues (Ghosh et al. 2021; Foulds et al. 2020). There might be an internal trade-off between fairness on the basis of attribute X_i and fairness on the basis of attribute X_j that cannot be removed or solved.

Building a *MinMax fairness* solution for algorithmic bias based on Rawlsian conceptualization of justice overcomes all these problems simultaneously, as we elaborate below.

(1.M) From Rawlsian *Theory of Justice* (Rawls 1971) we can extrapolate a clear fairness definition based on the *maxmin principle*, linked to the idea that *'A model is fair if it does not make more systematic errors for any sub-group in the dataset compared to the others'* (see Sect. 4, Definition 4.1 for *MinMax fairness* definition and the technical discussion about it). Ensuring fairness based on *maxmin principle* aims to increase the utility of any given model for those individuals who are mostly disadvantaged. Maximizing the minimum utility is equivalent to minimizing the maximum error within the context of predictive AI-algorithms, from which the name *MinMax fairness* is derived. This is possible via an elaborate but still simple justification applicable in a general way to all contexts, since the definition is linked to the errors made by the model. None of the other definitions of fairness does have this scope of applicability with such a clear justification that emanates from a rigorous connection with justice and equality of opportunities.

(2.M) *MinMax fairness* focuses on potential *systematic errors* that any given model could make in order to determine groups of disadvantaged people. In this respect, Rawlsian approach does not need to acquire the exact demographic information of individuals. Research confirms that those data points that suffer from systematic errors are most likely to represent people with sensitive characteristics that generate some form of disadvantage: see for example (Chow 1970; Varshney 2011; Kamiran et al. 2012) and the illustrative case in Sect. 4.2 of the present paper. The *MinMax fairness* solution introduced in this paper—which links the definition of disadvantages to systematic errors—attains a robust and generic definition that encompasses all combinations of potential known disadvantages. Consequently, in order to ensure algorithmic fairness towards disadvantaged people, a solution based on Rawlsian principles only depends on the errors made by the model and does not need to have the exact demographic characteristics of these people, solving one of the fundamental problems of algorithmic fairness (Andrus et al. 2021).

(3.M) Fairness based on *maxmin principle* aims to increase the utility for specific groups rather than ensuring parity: this does not create any abstract circumstance in which a decline in utility for a group is justified for the sake of reducing the difference between all groups. Rawlsian principles of fairness focuses on improving the absolute conditions of disadvantaged groups, rather than making their circumstances relatively less disadvantaged by pulling down the ones for the advantaged groups. An important aspect to consider when dealing with fairness interventions is the link between fairness and model accuracy. Indeed, alternative fairness interventions could have different impacts on the resulting model accuracy. As high-level intuition, while parity-based approaches are usually based on a trade-off between fairness and accuracy, the same does not necessarily apply to the proposed *MinMax fairness* solution, as this might result in *improving both fairness and model accuracy simultaneously*. Implementing this idea allows not only to help improving the utility for those groups who are known to be disadvantaged, but also to improve the utility for people suffering from systematic errors, who became disadvantaged within the context of a particular model and specific data circumstances. This is a fundamental benefit that not all fairness definitions allow to reach. It is important to stress the generality of such benefit, given the unforeseeable nature of all possible ways in which data patterns may single out hitherto unknown combination of characteristics as a source of disadvantage. At this stage, Rawlsian concept of *veil of ignorance* comes again into play in order to solve a problem usually under-stated for algorithmic bias. Disadvantages might be generated for specific groups as model-domain specific, and not necessarily confined to categories that are 'known'.

(4.M) Intersectionality (Ghosh et al. 2021) is an important element to take into account: when considering parity-based, it is possible that we ensure fairness on the basis of a specific 'known' attribute, and at the same time we make our model unfair w.r.t. another attribute unintentionally. This is not the case with *MinMax fairness*, as demographics are not needed to define fairness. In a way, intersectionality issues are 'not an issue', as potential mistakes for all protected groups are addressed simultaneously (including those that are not known or not acknowledged).

4 *MinMax fairness* as solution to algorithmic bias

This section proposes an explanation of the mathematical conceptualization of *MinMax fairness* solution to algorithmic bias by considering Rawlsian principles derived from his seminal contribution in the field of societal justice. We present and discuss why and how *MinMax fairness* provides a solution to overcome the current open challenges and deficiencies highlighted for parity-based approaches (Sect. 3.3), by leveraging on a specific mathematical setting to achieve fairness. Section 4.1 provides the definition of fairness based on *MinMax fairness* idea and describes the conceptual framework for its implementation; Sect. 4.2 discusses an illustrative example.

4.1 *MinMax fairness*

From a modelling perspective, one important open challenge is defining a solution on how to prevent the case in which the model might create potential discrimination towards specific groups of individuals on the basis of sensitive characteristics. Empirical evidence shows that *Fairness through unawareness* does not solve the problem of building fair models: often times, proxies for specific variables or indirect effects due to correlation between features could bring a certain level of bias into the modelling approach anyway (Pedreschi et al. 2008; Kusner 2017; Srivastava et al. 2019). Parity-based approach requires to specify a fairness definition, to identify specific sensitive attributes and to set a parity threshold. After de-biasing, the model performance often declines and unknown protected groups may suffer from the intervention.

Typically models might make large and/or more frequent systematic errors for vulnerable groups defined by various combination of protected attributes. Implementing a *MinMax fairness* solution implies to 'force' models to pay more attention to these type of errors, by introducing specific weights[3] for these errors (and associated datapoints) that can enable their identification. Intuitively, the weights have the goal of 'amplifying' the effect, and thus let the model directly pay more attention to these data points that suffer from larger/systematic errors. In the end, this will enable to help all vulnerable groups at the same time, by minimizing the maximum errors detected. The idea is that the model should ideally be equally well performing for all sub-groups, not making larger or more frequent errors for specific sub-groups. This conceptualization can be translated into the following *MinMax fairness* definition.

Definition 4.1 (*MinMax fairness*) A model is fair if it does not make more systematic errors for any sub-group in the dataset compared to the others.

[3] A discussion on the full detailed implementation of this technical solution based on a specific weighting function is beyond the scope of the present paper and is currently part of an on-going research project.

Table 1 Setting

ID	Observable features				Target value	Latent features	
	Received an education loan	Having permanent contract	Ever traveled abroad	Living in a suburb	Default history	Ethnicity	Disability
	X_1	X_2	X_3	X_4	Y	S_1	S_2
	$X_{1,j} \in \{0,1\}$	$X_{2,j} \in \{0,1\}$	$X_{3,j} \in \{0,1\}$	$X_{4,j} \in \{0,1\}$	$Y_j \in \{0,1\}$	$S_{1,j} \in \{A,B\}$	$S_{2,j} \in \{0,1\}$
	No, $X_{1,j} = 0$	No, $X_{2,j} = 0$	No, $X_{3,j} = 0$	No, $X_{4,j} = 0$	No, $Y_j = 0$		No, $S_{2,j} = 0$
	Yes, $X_{1,j} = 1$	Yes, $X_{2,j} = 1$	Yes, $X_{3,j} = 1$	Yes, $X_{4,j} = 1$	Yes, $Y_j = 1$		Yes, $S_{2,j} = 1$
i_1	No	No	Yes	No	No	A	No
i_2	Yes	Yes	Yes	No	No	A	No
i_3	No	Yes	Yes	No	No	A	Yes
i_4	Yes	No	Yes	No	No	A	No
i_5	No	Yes	Yes	No	No	A	No
i_6	Yes	No	No	Yes	No	B	Yes
i_7	No	No	No	Yes	Yes	B	No
i_8	No	No	No	Yes	Yes	B	No
i_9	No	No	No	Yes	Yes	B	Yes
i_{10}	No	No	No	Yes	Yes	B	No

The table reports a snapshot of the setting representing the basis of our heuristic example. We assume to have $n = 10$ people in the dataset, reported in column 'ID' as $\{i_1, i_1, \ldots, i_{10}\}$; for each individual i_j we assume to have the information regarding four binary 'Observable Features' $\{X_1, X_2, X_3, X_4\}$, two 'Latent Features' $\{S_1, S_2\}$ and one binary 'Target Value' Y_j

The concept of bias is linked to the errors made by the model, which result from potential false generalization (as discussed in Sect. 3.2, Figure 1). The operationalization of *MinMax fairness* definition can be described as follows:

- Bias is defined as function of the errors made by the model.
- Errors result from false generalizations, as for the data points belonging to the grey areas depicted in Figure 1: i) False generalizations may affect vulnerable groups defined by protected attributes, e.g. false 'bad' tendency; ii) False generalizations may affect privileged groups defined by protected attributes, e.g. false 'good' tendency.
- *MinMax fairness* solution attributes specific weights to errors resulting from false generalisation and then minimizes the maximum errors made by the model. The underlying idea is to identify in the model space the groups for which the model makes the largest errors and then 'force' it to perform better for these groups, thus reducing the errors.

Compared to other scientific contributions and practitioners' approaches to fairness, this solution has the novelty of being implementable without the need to know the sensitive attributes, by leveraging on the Rawlsian *veil of ignorance* principle. In its essence the idea is to "force" algorithms to pay more attention to those data points that suffer more from errors (both in terms of higher magnitude and being non-random) since these data points represent vulnerable groups of individuals. For a generic classification problem, from a technical point of view, this is possible by defining specific error and weight functions and optimize for the classifier in order to achieve the minimization of the maximum error. As a result, the model performance will improve. In line with Pareto optimality principles, this solution enables to achieve fairness by also reducing the disadvantage deriving from errors for the groups identified in the sample. From an implementation perspective, *maxmin principle* is also the basic idea of post-processing fairness interventions like (Kim et al. 2019) and blind Pareto fairness approaches (Martinez et al. 1970). Providing full details on the mathematical formulation is beyond the scope of the present paper, whose aim is to provide the high-level intuition for the foundations and is left to a dedicated technical research exploration which is currently on-going. However, to provide insights about the practical issues related to this solution, Sect. 4.2 discusses one illustrative example of *MinMax fairness* implementation.

4.2 Illustrative example

In order to show how *MinMax fairness* provides a solution to algorithmic bias, we introduce the following illustrative example within the context of credit decision modeling (i.e. loan applications). It is important to emphasize that this is meant to be a heuristic example serving the purpose of providing a setting to discuss different fairness interventions; as such, we have developed it by deliberately avoiding or simplifying technical aspects.

4.2.1 The dataset

Let us assume to have a dataset consisting of $n = 10$ people, and assume they all received loans in the past. Table 1 reports a snapshot of the setting we have, including binary 'Observable Features' $\{X_1, X_2, X_3, X_4\}$, binary 'Target Value' Y and 'Latent Features' $\{S_1, S_2\}$. For each individual i_j, the dataset contains:

- The information associated to four binary characteristics given as 'Observable Features' $\{X_1, X_2, X_3, X_4\}$, namely i) X_1 : whether they ever received a loan to finance their education, ii) X_2 : whether they have a permanent contract, iii) X_3 : whether they have ever travelled abroad in the past, and iv) X_4 : whether they live in a suburb. For all these features we consider 'no' associated with 0 and 'yes' associated with 1.
- The information regarding the 'Target Value', meaning the 'default history'; for the 10 individuals we know whether they 'defaulted' or paid back ('no default'), i.e. column with binary feature 'default history'. We denote with $Y_j \in \{0, 1\}$ the value corresponding to 'no', meaning 'no default' ($Y_j = 0$) or 'yes', meaning 'default' ($Y_j = 1$).
- The information associated to two 'Latent Features' $\{S_1, S_2\}$, that are linked to sensitive characteristics, i.e. ethnicity and disability. We assume that: i) we can access the information about ethnicity but avoid using it in our model deliberately; ii) we have no access to disability information. Independently of the specific case, both 'Latent Features' are part of the reality.

In this respect, in general terms, the following holds:

1) We assume to have two alternative values for 'ethnicity', namely A and B. In this dataset, all people with ethnicity B live in suburban areas. We do not want to use ethnicity as a factor in our decision making process; at the same time, ethnicity information can enter into our model unintentionally through the information provided by residence (i.e. suburban dwelling) captured by X_4.
2) In this dataset there are three persons with disability (individuals i_3, i_6, i_9) and interestingly one of these persons, individual i_6, is among the three people who received a loan during education; at the same time, this is also the only person in the dataset with this unique combination (i.e. having disability and having had a loan during education).

Despite being a highly stylized example, this is a dataset which we might use for developing an algorithm/model to facilitate decision making regarding accepting or rejecting loan applications. To build such a model, we need to use the information provided by these 10 people on the 'Observable Features' and on the 'Target Value' (whether they defaulted or not). In this way we can predict whether the next person that would apply for a loan would pay back or default. This is a classical setting in which we may want to check if our model is fair and correct it, in case it is not.

4.2.2 The model

Let us suppose to consider a simple linear probability model[4], in which each 'Observable Feature' $\{X_1, X_2, X_3, X_4\}$ is scaled by a coefficient $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ and the resulting sum is an estimate of the default probability, that we denote as PD. We formalize a generic model M as follows:

$$M: \quad \alpha_1 \cdot X_1 + \alpha_2 \cdot X_2 + \alpha_3 \cdot X_3 + \alpha_4 \cdot X_4 = PD, \tag{1}$$

meaning that we aim to find how each of the 'Observable Features' is contributing to the prediction of the default probability. To train the model, we should consider all individuals in the dataset. As an example, for individual i_1 we have:

$$\alpha_1 \cdot 0 + \alpha_2 \cdot 0 + \alpha_3 \cdot 1 + \alpha_4 \cdot 0, \tag{2}$$

and the aim of training the model is to find the optimal set of coefficients that minimize the errors made by the model. By training the generic model in Eq.(1) on the dataset provided in Table 1 we obtain the following coefficients:

$$-0.407 \cdot X_1 - 0.068 \cdot X_2 + 0.203 \cdot X_3 + 0.881 \cdot X_4 = PD, \tag{3}$$

to estimate the model's prediction on the PD. The results of Eq.(3) are real values between 0 and 1, representing the default probability[5] estimated for each individual $i_j, j \in \{1, \ldots, n\}$ in the sample. From the real values predictions PD_j, we can go back to qualitative predictions by introducing a threshold[6]. Let us consider $t = 0.4$ as threshold, and

[4] We observe that Linear Probability Models (LPM) can sometimes produce unrealistic predictions $p < 0$ or $p > 1$; in such cases, one should consider predictions $p < 0$ as 0 and predictions $p > 1$ as 1. There are also more sophisticated solutions (Mood 2010, p.81). In some circumstances, LPM can actually be more suitable than logistic regression for probability modelling (see (Mood 2010, p.78) and (Caudill 1988)). Here we use this model for its simplicity and due to its suitability for the purpose of a heuristic example.

[5] Depending on the type of model used, direct interventions might be needed to keep values between 0 and 1. In this case, for Linear Probability Models (LPM), see also footnote 3.

[6] This is typically an expert-based parameter.

Table 2 Training the model: results

ID	Target value	Predicted probability	Predicted classification	Error
	$Y_j \in \{0, 1\}$	$PD_j \in [0, 1]$	$PD_j \geq 0.4$	$e_j \in [0, 1]$
	No, $Y_j = 0$			
	Yes, $Y_j = 1$			
i_1	No	0.203	No	0.203
i_2	No	0.000	No	0.000
i_3	No	0.136	No	0.136
i_4	No	0.000	No	0.000
i_5	No	0.136	No	0.136
i_6	No	0.475	Yes	0.475
i_7	Yes	0.881	Yes	0.119
i_8	Yes	0.881	Yes	0.119
i_9	Yes	0.881	Yes	0.119
i_{10}	Yes	0.881	Yes	0.119

The table reports the results from the training phase on the model given in Eq. (1). In particular, the table contains: the ‘Target Value’ (provided in Table 1), the predicted probability of default (PD_j estimated via Eq. (3)), the predicted classification depending on the threshold $t = 0.4$ and the errors made by the model (computed based on Eq. (4))

the following rule for the classification of creditworthiness: any probability exceeding 0.4 (i.e. $PD_j \geq 0.4$) is an indication of high risk of default and corresponds to the prediction “individual i_j would have defaulted”. Like all models, this one too makes errors; what matters is that we aim to have a model that gets close enough to 0 or 1 in terms of predicted default probability in order to make the correct classification. We estimate the error e_j for each individual as the absolute value of the difference between the estimated default probability and the ‘Target Value’, namely

$$e_j = |PD_j - Y_j|, \tag{4}$$

with Y_j being the ‘Target Value’ for individual i_j. Table 2 shows the outcomes of the training phase and the errors made by the model.

4.2.3 Fairness through intervention: parity-based vs *MinMax fairness*

Looking at the outcomes of the model, we observe that, apart from individuals i_2 and i_4, all other individuals suffer from errors. At the same time, we also observe that individual i_6 is suffering from the largest error so that s/he is actually classified wrongly. By ranking the errors, individual i_6 is followed by individual i_1, though this person—despite the second biggest error—is not wrongly classified.

It is important to notice that although we do not use ethnicity as a feature in our model, it still has a crucial influence on the outcome of the model via X_4 as a proxy. Thus, belonging to ethnicity B (captured by living in suburban area through X_4) increases the probability of default due to its large positive coefficient 0.881). It is quite possible not to be aware of such proxies and indirectly use sensitive attributes in the model. This example also shows why *Fairness through unawareness* is not a reliable way of addressing algorithmic bias. Thus, the focus below is on *Fairness through intervention* in the form of *parity-based* and *MinMax fairness*.

Parity-based fairness. Let us now consider how fairness enters into play, and how to make interventions, starting from the parity-based approach. For this purpose, we consider demographic parity (Feldman et al. 2015), as it is one of the most commonly used parity-based fairness definitions. This approach requires to ensure that a proportion of the desired predictions (such as getting the loan request approved) is sufficiently similar across privileged and unprivileged groups defined by a protected attribute. Here the rule of thumb emanating from the inheritance coming from the US Labor legislation[7] is to attain 80% parity among groups.

For illustrative purposes, let us assume that we have access to ethnicity but we do not have access to disability. Under these conditions, we may choose to proceed with this knowledge and make sure that the proportion of desired outcomes (e.g. ‘no default’) would be identical for individuals belonging to ‘group A’ (ethnicity A, i_1, i_2, i_3, i_4, i_5) and ‘group B’ (ethnicity B, $i_6, i_7, i_8, i_9, i_{10}$). From Table 2 one can see that the model makes 5/5 desired outcomes for ‘group A’ while 0/5 for ‘group B’ (column ‘Predicted Classification’). To ensure demographic parity, we would need to intervene by changing the predictions for ‘group B’ in order to have 4/5 “no default” predictions. By doing this, what can happen is that we would probably correct the wrong prediction for individual i_6 but then predict that three persons—who would actually have defaulted—would pay their loans. As a result, this would create adverse conditions for new applicants as it would increase the likelihood of the model to make wrong predictions for those who would default by evaluating them as creditworthy, possibly creating dire financial circumstances for them. In fact, if the goal is to ensure that people would not be made vulnerable and discriminated by the model (and due to the model), then demographic parity would not serve this purpose: while possibly eliminating the vulnerability of individual i_6, it would generate new vulnerabilities for other individuals by decreasing the model’s accuracy (discussion point 2.P, Sect. 3.3).

[7] The legal origin of the 80% rule is a particular US legislation (US 1964) but it has been explicitly articulated in another document (U.S. Equal Employment Opportunity Commission (EEOC) 1978).

Another important point: since we do not have access to disability, the parity that we would attain based on ethnicity would not necessarily ensure fairness based on disability; this could only happen by chance: for the three persons with disability (i.e. individuals i_3, i_6, i_9 from Table 1) the proportion of desired outcomes is 1/3 while the corresponding proportion for the remaining people without disability is 4/7. We know now that imposing ethnicity-based fairness would require giving "no default" predictions to four individuals belonging to 'group B'; if one of these people is, by chance, disabled, then we can attain a parity of 85% between disabled and not disabled people. There are five candidates (individuals $i_6, i_7, i_8, i_9, i_{10}$) and only two of them with disability (individuals i_6, i_9): thus, the probability of attaining parity on the basis of disability while attaining parity on the basis of ethnicity is just 0.40, meaning that there is only 40% chance of accomplishing that (discussion point 3.P, Sect. 3.3). In the remaining cases, parity-based intervention will create a different form of unfairness, e.g. towards a different group of people.

Obviously, looking at parity-based approaches, we could also choose another parity-based fairness definition (for instance parity in terms of proportion of wrong predictions): in such a case, we would not only encounter the exact same problems discussed for demographic-parity (creating new vulnerabilities, leaving fairness on the basis of disability to chance, reducing model's accuracy), but we would also need to justify our choice of fairness definition at least for those people who argue that demographic-parity would be better. Moreover, due to intersectionality issues (Ghosh et al. 2021; Foulds et al. 2020), it might be that ensuring fairness on the basis of a specific attribute (e.g. ethnicity) generates unintentionally discrimination on the basis of another one (e.g. disability, see also discussion points 1.P, 4.P Sect. 3.3).

MinMax fairness. Let us now consider *MinMax fairness*. Based on Definition 4.1, the premise is that fairness implies to make sure that the model increases the utility associated to disadvantaged individuals, and we measure the utility of the model by looking at the errors it makes (discussion points 1.M-2.M, Sect. 3.3). Utility decreases with the magnitude of the errors, i.e. the smaller the error, the higher the utility. From this perspective, it is clear that the only person that suffers from bias is individual i_6 (discussion point 3.M, Sect. 3.3): while s/he has not defaulted, the model makes a very large error and puts the prediction above the threshold and classifies this person as non creditworthy, i.e. 'default' (Table 3). The implication is that the resulting model will make such wrong predictions *systematically* for new applicants with similar characteristics.

It is useful to shortly reflect on why the model is making this large mistake: in fact, this is due to a false generalisation. Individual i_6 resembles to four individuals (i_7, i_8, i_9, i_{10}) who have defaulted: like them, i_6 lives in a suburb, has no permanent contract, has never travelled abroad. However, there is one difference, related to i_6 having received a loan during her/his education. Apparently, this is the only 'observable' difference. The reason might be hidden in the 'Latent Features': perhaps, because of her/his disability, s/he needed extra means for education and developed an attitude to acquire and repay loans. Obviously, one can think of many other scenarios. The evidence which remains is that there is one observable difference between individual i_6 and four individuals who have defaulted: in this case, having received a credit during education (X_1). Based on how the model works, it gives more attention to similarities, rather than this difference and therefore makes a very large error, generating a wrong predicted classification. In order to prevent this, we need to 'force' the model to pay more attention to what makes i_6 distinguishable. Technically, this can be accomplished by re-weighting the dataset and assign a higher weight to individual i_6 compared to the others. Let us suppose to assign to individual i_6 a weight bigger than the one assigned to all other individuals: conceptually, this is equivalent to 'forcing' the model to put more attention to this data point, similarly to use a magnifying glass so that the model can 'see' it better.

Table 3 Training the model: results

ID	Target value	Predicted probability	Predicted classification	Error
	$Y_j \in \{0, 1\}$	$PD_j \in [0, 1]$	$PD_j \geq 0.4$	$e_j \in [0, 1]$
	No, $Y_j = 0$			
	Yes, $Y_j = 1$			
i_1	No	0.267	No	0.267
i_2	No	0.000	No	0.000
i_3	No	0.178	No	0.178
i_4	No	0.000	No	0.000
i_5	No	0.178	No	0.178
i_6	No	0.311	No	0.311
i_7	Yes	0.844	Yes	0.156
i_8	Yes	0.844	Yes	0.156
i_9	Yes	0.844	Yes	0.156
i_{10}	Yes	0.844	Yes	0.156

The table reports the results from the training phase on the model given in Eq. (1). In particular, the table contains: the 'Target Value' (provided in Table 1), the predicted probability of default (PD_j estimated via Eq. (3)), the predicted classification depending on the threshold $t = 0.4$ and the errors made by the model (computed based on Eq. (4))

When we estimate the model including the weights[8] we obtain the following estimates:

[8] The example considers a weight $w_6 = 2$ for individual i_6 and a weight equal to 1 for all the others.

$$-0.533 \cdot X_1 - 0.089 \cdot X_2 + 0.267 \cdot X_3 + 0.844 \cdot X_4 = PD. \tag{5}$$

The estimates based on Eq.(5) produce the errors and predictions given in Table 3. This new model also makes errors but what matters is that none of these errors is large enough to lead to a wrong creditworthiness prediction (column 'Predicted Classification'). The error for individual i_6 declined from 0.475 to 0.311 and thanks to this intervention the classification is now correct: 'no default'. With this set of coefficients, if the model is used to make predictions, it would help people who suffer from vulnerability exactly as it did for individual i_6.

Overall, it is important to stress the advantages of this approach: no arbitrary fairness definition; no sensitive attribute is needed to identify groups and impose fairness; the method helped the vulnerable person in this dataset without creating new vulnerabilities for this person and/or other people, as this solution helps protected groups simultaneously; the method has not reduced the model accuracy, and instead increased it (discussion points 3.M-4.M, Sect. 3.3).

5 Conclusion

The aim of this paper is to show how Rawlsian *Theory of Justice* (Rawls 1971) can help to solve some of the current open challenges related to algorithmic fairness and to provide its epistemological motivation. The contribution of the paper is to introduce Rawlsian principles in their original form and substantiate why they can represent a solution to some technical open questions for algorithmic fairness. To illustrate this, the paper provides an overview of the main Rawlsian ideas and principles, discusses their implementation and provides an intuition of the mathematical perspective for algorithmic bias, i.e. *MinMax fairness*. The paper introduces a high-level formalization of the solution in the context of modelling and algorithmic decision making and discusses the main benefits in achieving algorithmic fairness via *MinMax* when compared to other existing approaches. This is done both at conceptual level and via an illustrative example. In summary, when compared to parity-based approaches, *MinMax fairness* as solution to algorithmic bias shows at least four advantages: 1) there is no need to make a context-dependent fairness definition subject to arbitrary choices; 2) knowledge of the protected attribute information to implement the fairness solution is not needed; 3) it does not aim to ensure parity of outcomes between groups, and rather to increase the utility for specific groups—this does not happen at the cost of reducing utility for other groups; as a result, the accuracy of the model for unprivileged groups might improve without reducing the one for the privileged groups, overcoming the fairness-accuracy trade-off; 4) intersectionality issues (Ghosh et al. 2021) are 'not an issue', as potential mistakes for all protected groups are addressed simultaneously (including those that are not known or not acknowledged). The discussion focuses not only on pure engineering choices but also on the overall ethical perspective in the approach to models and their use for our society. The present paper proposes *MinMax fairness* as solution for algorithmic bias in the form of a fairness intervention; from this perspective, implementing *MinMax fairness* enables to go towards fairer and more accurate models, overcoming the classical trade-off between increasing fairness and reducing model performance faced when implementing parity-based approaches. It is crucial to point out that structural problems of societies cannot be solved by changing the algorithmic fairness approach. What parity-based and *MinMax fairness* do not solve—and cannot solve—are the structural and intrinsic justice problems of our society present in different domains (e.g. banking, health, welfare, etc.): no algorithm alone or fairness intervention can solve societal issues or ameliorate injustices that are structural. However, as we show in this paper, *MinMax fairness* can at least prevent the deepening and expanding of existing injustices that may be historically and structurally part of our societies.

Acknowledgements Disclaimer: The opinions expressed in this paper are solely those of the authors and do not necessarily represent those of their current or past employers.

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

References

Aler Tubella A, Barsotti F, Kocer RG, Mendez J (2022) Ethical implications of fairness interventions: what might be hidden behind engineering choices? Ethics Inf Technol 24:12

Andrus M, Spitzer E, Brown J, Xiang A (2021) What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 249–260

Barry B (2017) Equal opportunity and moral arbitrariness. Distributive justice. Routledge, Amsterdam, pp 213–234

Bellamy RKE, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D et al (2019) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM J Res Dev 63(45):4:1-4:15

Caudill SB (1988) An advantage of the linear probability model over probit or logit. Oxford Bull Econ Stat 50(4):425–427

Chow C (1970) On optimum recognition error and reject tradeoff. IEEE Trans Inf Theory 16(1):41–46

Clark GL (1993) Making moral landscapes: John Rawls' original position. Polit Geogr Q 5(4):147–162

Daniels N (2003) Democratic equality: Rawls's complex egalitarianism. In: Samuel RF (ed) The Cambridge companion to Rawls. Cambridge University Press, Cambridge, pp 241–276

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. ITCS 2012—innovations in theoretical computer science conference. ACM Press, New York, pp 214–226

Dworkin R (1973) The original position. Univ Chicago Law Rev 40(3):500–533

EBA (2020) Report on big data and advanced analytics. Technical report, European Banking Authority

European Commission (2018) Reform of EU data protection rules

European Commission (2019) Ethics guidelines for trustworthy AI. Technical report

European Commission (2021) European Act on the use of AI

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 259–268

Fleurbaey M (1995) Equal opportunity or equal social outcome? Econ Philos 11(1):25–55

Foulds JR, Islam R, Keya KN, Pan S (2020) An intersectional definition of fairness. In: IEEE 36th international conference on data engineering (ICDE). Association for Information Systems, pp 1918–1921

Freeman S (2009) Justice and the social contract: essays on Rawlsian political philosophy. Oxford University Press, Oxford

Ghosh A, Genuit L, Reagan M (2021) Characterizing intersectional group fairness with worst-case comparisons. In: Proceedings of machine learning research, 2nd workshop on diversity in artificial intelligence (AIDBEI), vol 142. Association for Information Systems, pp 22–34

Haas C (2020) The price of fairness—a framework to explore trade-offs in algorithmic fairness. In: 40th international conference on information systems, ICIS 2019. Association for Information Systems

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3323–3331

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. In: International conference on AI algorithm, pp 1929–1938

Joseph M, Kearns MJ, Morgenstern JH, Roth A (2016) Fairness in learning: classic and contextual bandits. In: NIPS

Kamiran F, Karim A, Zhang X (2012) Decision theory for discrimination-aware classification. In: 2012 IEEE 12th international conference on data mining, pp 924–929

Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: International conference on machine learning. PMLR, pp 2564–2572

Kim M, Ghorbani A, Zou J (2019) Multiaccuracy: black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. Adv Neural Inf Process Syst 30:25

Kymlicka W (2002) Contemporary political philosophy: an introduction. Oxford University Press, Oxford

Lahoti P, Beutel A, Chen J, Lee K, Prost F, Thain N, Chi E (2020) Fairness without demographics through adversarially reweighted learning. Adv Neural Inf Process Syst 33:728–740

Martinez NL, Bertran MA, Papadaki A, Rodrigues M, Sapiro G (1970) Blind Pareto fairness and subgroup robustness. In: International conference on AI algorithm, pp 7492–7501

Mood C (2010) Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur Sociol Rev 26(1):67–82

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 560–568

Rawls J (1971) A theory of justice. Harvard University Press, Harvard

Rawls J (1985) Justice as fairness. Philos Rev 67(2):164–194

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

Rawls J (2001) Justice as fairness: a restatement. Harvard University Press, Harvard

Roemer JE (2002) Egalitarianism against the veil of ignorance. J Philos 99(4):167–184

Sarah B, Miro D, Richard E, Brandon H, Roman L, Vanessa M, Mehrnoosh S, Hanna W , Kathleen W (M2020) Fairlearn: a toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft

Srivastava M, Heidari H, Krause A (2019) Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining

U.S. (1964) Equal Employment Opportunity Commission (EEOC). Title VII of the Civil Rights Act of 1964. In United States Code, vol 42, pp 88–352

U.S. Equal Employment Opportunity Commission (EEOC) (1978) Uniform guidelines on employee selection procedures—part 1607. In United States Code (Title 29, Labor. Subtitle B), vol 4, pp 88–352

Varshney KR (2011) A risk bound for ensemble classification with a reject option. In: 2011 IEEE statistical signal processing workshop (SSP), pp 769–772

Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM international workshop on software fairness (fairware), pp 1–7

Yang CS, Dobbie W (2020) Equal protection under algorithms: a new statistical and legal framework. Mich Law Rev 119(2):291–395

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