Entitlement Justice and Measures of Algorithmic Fairness

Edward Speer
California Institute of Technology
espeer@caltech.edu

Llama-2-7b
Meta Platforms, Inc.

January, 2025

Abstract

This paper explores the relationship between entitlement justice and measures of algorithmic fairness. ...

1 Introduction

The rise of algorithmic decision making in the public sector has caused significant public concer. As algorithms increasingly make decisions that affect individials' lives, from determining creditworthiness to predicting criminal recidivism, the public has grown fearful of their potential to perpetuate existing social inequalities. A 2018 study showed that 58% of Americans believe that algorithms will always have some level of bias Smith (2018), and as we know from the famed COMPAS case, these fears are not unfounded Angwin et al. (2016).

In response to these concerns, researchers have developed two broad and increasingly vast bodies of work. The first, which we will refer to as *algorithmic fairness*, focuses on developing statistical and computational tools to ensure that algorithms do not discriminate against protected groups. The second, referred to as *algorithmic accountability*, focuses on explainability and interpretability — developing tools to help users understand and interpret the decisions made by algorithms. The former area of research is what we will focus on in this paper.

The field of algorithmic fariness is often conceptualized as the application of the philosophical notion of distributive justice to algorithmic decision making. At first glance, this seems like a natural fit. The goal of distributive justice is to ensure that the allocation of the benefits and burdens of society are distributed fairly among its members, and the goal of algorithmic fairness measures are to ensure that the allocation of decisions by algorithms complies with some notion of fariness. However, recent work questions this analogy Hertweck et al. (2024), analytically showing that the extent to which algorithmic fairness measures can be seen as a form of distributive justice is quite limited, and isolated to egalitarian concepts of justice Kuppler et al. (2021).

In this paper, we propose a new direction for research that incorporates a previously overlooked distributive justice concept: entitlement justice. Entitlement theory, which roots justice in the idea of respecting individuals' property rights, offers a more nuances and context-sensitive understanding of algorithmic fariness. We argue that by incorporating entitlement justice into the

design of algorithmic fairness measures, we can create a more robust framework for evaluating algorithmic decisions. When this framework is applied to the broader sociotechnical systems in which algorithms are embedded, we can better understand the social implications of algorithmic decision making and develop more effective strategies for mitigating their negative effects.

The rest of this paper is organized as follows. In Section 2, we provide an overview of the existing literature on algorithmic fairness and distributive justice. We draw on the formalism from Kuppler et al. (2021) and Corbett-Davies et al. (2023) to create a unified model for understanding algorithmic fairness and distributive justice consistently with each other. In Section 3, we introduce the concept of entitlement justice and discuss its historical development. We confront the traditional objections to entitlement theory and show how they can be overcome in the context of algorithmic decision making. In Section 4, we propose a new framework for understanding algorithmic fairness through the lens of entitlement justice. We analyze the implications of this framework for existing algorithmic fairness measures and show an example of how it can be applied to a real-world case study. Finally, in Section 5, we conclude with a discussion of the broader implications of our work and suggest directions for future research.

2 Background

Within this paper, we will restrict our attention to cases where the decision problem under scrutiny can be formulated using the following very simple formalism. Some entity (algorithmic or otherwise) possesses a finite amount of resource X, and must allocate it among a set of agents $A = \{a_1, a_2, \ldots, a_n\}$. Considerable work is done here by the term *resource*, which we will define simply as an element which may be distributed among agents. Traditional examples include money and admissions, but we will also consider more abstract resources such as representation or influence. Following Kuppler et al. (2021), we will define a *distribution rule* as a statement in the form: Allocate amount R of resource X to agent a_n iff X has attribute Y. Broadly speaking, the role of a theory of justice is now to provide a justificable and explainable Y for a given distribution rule, while the role of an ideal algorithmic fairness measure is to evaluate the extent to which an algorithmic-decision maker obeys a particular distribution rule.

Admittedly, not all problem domains in which algorithmic decision-making is applied can be formulated in this way Green and Hu (2018). For example, applications of AI in natural language translation may not be easily formulated in terms of resource allocation, but the reader may still be concerned with the perpetuation of social biases through the decisions it makes in translation. While these cases are significant, they do not fall simply within the domain of distributive justice, and so we will not consider them here.

2.1 Algorithmic Fairness Measures

In the canonical presentation of algorithmic fairness, we are given a population of agents $A = \{a_1, a_2, \ldots, a_n\}$ with observed covariates X drawn from some distribution $P(X)$. We are told that some set A of protected attributes may be derived from X. Each agent a_n in the population is subjected to a binary decision according to some decision rule $d : X \to \{0, 1\}$ Corbett-Davies et al. (2023).

In the formal setting we described above, this decision rule plays a clear role. There is some distribution rule which we would like to enforce in the allocation of resources, such that each a_n receives an amount of resource X according to some attribute y. The decision rule d attempts to determine the "y-*ness*" of each agent a_n based on the observed covariates for that agent x_n. y is highly unlikely to be directly observable or straightforward, and we are unlikely to be able to predict it perfectly from the information delivered by x_n. The decision rule d is then a function which imperfectly approximates the desired distribution rule, making errors at some frequency. Algorithmic fairness measures presented in the literature thus may be seen as defining the following attributes of the decision rule d:

- The relationship of the attribute y to the set of protected characteristics A
- The method of detection of errors made by a decision rule d

Some of the most commonly discussed measures in the literature are presented below together with their critiques leveraging this lens. For a more exhaustive list of measures, see Corbett-Davies et al. (2023).

Definition 1. *Demographic Parity — A decision rule d is said to satisfy demographic parity if the probability of receiving a positive decision is independent of the protected attributes A Dwork et al. (2012).*

Demographic parity asserts the condition that the probability of predicting that an agent having the attribute y cannot depend on the protected attributes A. When using demographic parity as a fairness measure, therefore we measure errors made by the decision rule d by the extent to which the probability of receiving a positive decision depends on the protected attributes in A.

From this presentation it is easy to see multiple ways in which demographic parity may be unsatisfactory. For example, our decision rule d is allowed to be very poor at predicting whether agents have the attribute y, so long as it is equally poor across all protected attributes. Mistakes in predicting y are not penalized, and so could be patterned in a way that is harmful to some protected groups. For example, d could produce many false positive predictions of y for one group and many false negative predictions for another as long as the resulting distribution is balanced correctly Barocas et al. (2017).

Definition 2. *Equalized Odds — A decision rule d satisfies equalized odds if the true positive rate and false positive rate do not vary with respect to A Hardt et al. (2016).*

Equalized odds may be thought of as again positing that attribute y may not depend on A, but it goes further to say that errors in our prediction of y — specifically false positives—must be distributed uniformly across groups. We thus measure errors as dependencies between the false positive rate and A.

Equalized odds is often critiqued for struggling to deal with unequal base rates between groups. Consider the following example from criminal justice. We have a distribution rule which says to allocate a parole to a prisoner if they are very unlikely to recidivate. Due to a history of discriminatory practices and social marginalization, black prisoners have a base rate of recidivism

much higher than white defendants Crime and Alliance (2023). As a result, allocating a parole to a white prisoner has a base line lower likelihood of being a false positive. Therefore, when we perform post-processing of our data to balance false positive rates, we may actually *add* false positives to the white portion of the dataset, resulting in an increase in the number of white prisoners receiving parole.

Definition 3. *Counterfactual Fairness — A decision rule d satisfies counterfactual fairness if protected attributes from A do not play a causal role in its output Kusner et al. (2018).*

Counterfactual fairness posits a different criteria about y than demographic parity or equalized odds. Rather than mandating that y should be independent from features in A, counterfactual fairness mandates that attributes in A cannot be causes of y. We therefore measure errors in the prediction of y by detecting causal links between A and the prediction of y.

Counterfactual fairness is often critiqued based on the difficulty and potential subjectivity of detecting causal links between variables. Recent work on the social construction of demographic variables reveals that causal modeling may have an inherently normative basis Hu (forthcoming), and even if these issues are set aside, the computational expense of causal discovery can create issues of practicality.

This discussion of dominant algorithmic fairness measures and their shortcomings motivates further discussion of theories of distributive justice. To what extent do the features of y posited by these measures align with the rules of distribution dictated by theories of algorithmic justice? Is it valid to say that these measures enforce distributive justice in any way? And it is possible to address or understand their shortcomings in terms of the philosophy of distributive justice?

2.2 Theories of Distributive Justice

The role of a theory of distributive justice is to provide the rules of distribution that define fairness in society. Specification of these rules is exactly the process of defining a y in the formalism we have presented. Several conventional theories of distributive justice have been proposed in the literature, and we will discuss a few of them here. For a more exhaustive list of theories presented in this framework, see Kuppler et al. (2021)

Definition 4. *Egalitarianism — Egalitarianism posits that all individuals should receive an equal share of resources Rawls (1971).*

Equally restated in the formalism we have presented, egalitarianism posits that amount R of resource X should be allocated to agent a_n if and only if the doing so minimizes the overall inequality across the population. Thus in this case, y is the property of *lacking* good X relative to the population.

Note that the precise currency of egalitarianism is unclear. For example, allocating money to an individual who is lacking in food only indirectly impacts the stock of concern, but it is clear that doing so will still reduce the overall inequality of the population. Clearly in this case we could shift our currency to general wealth or utility, but the choice of currency is not immediately obvious, nor does it seem generally possible to define a currency which is globally applicable Binns (2018).

Definition 5. *Sufficientarianism — Sufficientarianism posits that all individuals should receive a share of resources sufficient to meet some threshold level of well-being Sen (1979).*

Sufficientarianism is a theory of distributive justice which posits that the distribution rule should be such that all agents receive an amount of resource X which is sufficient to meet some threshold level of well-being. Thus y is defined as the property of *needing* good X. Note that what constitutes a threshold level of well-being and what goods are required for it is not immediately clear, and that under conditions of scarcity it may be impossible to meet the threshold for all agents.

Definition 6. *Desert — Moral desert is the idea that individuals should receive resources in proportion to their moral worth as measured by some metric of merit Pojman (1997).*

Theories of desert therefore set y to be some form of moral merit, and the distribution rule is such that agents receive X if they deserve it according to their merit. Theories of desert have been critiqued for being highly subjective and difficult to measure. Any attempt to craft a metric for moral merit is likely to be highly controversial and may be subject to manipulation by those in power.

This formulation lays bare the issues with considering algorithmic fairness measures as enforcing distributive justice. In most cases it is unclear how the features of y posited by these measures align with the rules of distribution dictated by theories of distributive justice. For example, demographic parity may appear to enforce some form of egalitarianism by ensuring that the output distribution of the decision rule is equal across protected groups. However, this is a failure in two ways. Firstly, egalitarianism mandates that allocations be balanced across all individuals, not across groups. Secondly, our measurement of errors in the decision rule d is based solely on the distribution enforced by d, not on the actual distribution of resources in society. In cases with a large pre-existing disparity, enforcing parity might be thought of as preventing the widening of the gap, but not as fully enforcing egalitarian justice.

Similar complaints may be made about the other fairness measures presented here, and it is clear that the relationship between algorithmic fairness and distributive justice is not straightforward. This motivates further discussion of the relationship between these two fields, and the potential for a more cautious and nuanced approach to the application of algorithmic fairness measures.

3 Entitlement Justice

An entitlement theory of justice is a distributive theory of justice which posits the following distribution rule: Allocate amount R of resource X to agent A if and only if A is entitled to R of X. An entitlement in this context is a *property right* held by the agent over the resource. Different entitlement theories of justice differ in the criteria they use to determine entitlements, and the concept of property rights they endorse. Here we will detail the entitlement theory of justice as proposed by Nozick (1974) and its issues, then discuss more recent efforts at reconciling the theory with the demands of justice.

3.1 Nozick's Entitlement Theory of Justice

Nozick's entitlement theory of justice, often called the concept of libertarian justice, is a theory of justice that was developed as a fundamental challenge to Rawl's liberal egalitarianism. On the liberal egalitarian view, ensuring justice is an inherently redistributive task. The justice of a distribution of resources is determined by the extent to which it is equal over individuals, and there is an implied moral responsibility to redistribute resources to those who lack them to increase the overall equality of the distribution. This ideology provides a strong defense of taxation and welfare programs, which redistribute resources in order to flatten the distribution of wealth Rawls (1971).

Libertarian justice takes issue with the consequences of adopting this view. Nozick asks us to consider a thought experiment. Suppose we began with an equal distribution of resources across society. People in this society have the freedom to choose how to use their resources, and to exchange them with others as they feel is fair. Many people are willing to pay to see Wilt Chamberlain play basketball, and so they each pay him a small amount of money to see him play. Over time, Chamberlain will accumulate a large sum of money though his efforts. The distribution of resources in the society will no longer be equal, but will be skewed towards Chamberlain. On the egalitarian account, this excess wealth that Chamberlain has accumulated is unjust, and must be taken and redistributed across society. On the libertarian view, however, Chamberlain has gained an entitlement to his accumulated wealth, and to take it away from him is akin to stealing. After all, if this wealth is taken away from him, then he will have received nothing for his efforts, and enjoyed no fruits of his labor.

In Nozick's theory, people gain entitlements over resources in accordance with 3 principles:

1. The principle of justice in acquisition: A person who acquires a resource through a just process is entitled to that resource. A process of acquisition is just if the acquisition is in accordance with Lockean proviso (discussed below).

2. The principle of justice in transfer: A person who acquires a resource through a just transfer is entitled to that resource. A transfer is just if the transfer is voluntary and the resource is transferred from someone who is entitled to it.

3. The principle of rectification: A person who acquires a resource through the rectification of a prior injustice is entitled to that resource. Rectification must be proportional to the injustice which is being rectified.

On analysis, one will see that a key difference between this libertarian view and the liberal egalitarian view is the fundamental unit of justice. For the liberal egalitarian, justice is realized in the distribution of resources itself. This approach is refered to as a patterned or end-state view of justice. For the libertarian, justice is realized in the process by which resources are acquired and transferred. This approach is refered to as a historical theory of justice. In order to determine if the current state of affairs is just with respect to a particular holding, one must trace the history of that holding back to its original acquisition, and ensure that each step in the process was just. For Nozick, any end-state theory of justice is inherently flawed, as it requires the restriction of individual liberties Henberg (1977). It is plain that this view of justice hinges strongly on being

able to identify and justify the initial acquisition of resources, else the theory can say nothing about the justice of the current distribution of resources.

3.2 The Lockean Proviso

For Nozick the Lockean proviso underscored the principle of justice in acquisition. The proviso contains two parts. The first part is a mechanism for justifying the initial acquisition of resources. It begins with the inherent right of self-ownership that all individuals possess. Locke argued that when an individual mized their own labor with a resource, they transferred some of themself into the resource, and so extended their right of self-ownership over the resource, thereby obtaining an entitlement to it. The second part of the proviso, almost as an afterthought, is a restriction on the extent to which resources can be acquired. It states that a person can only acquire a resource if there is enough and as good left over for others. This restriction is necessary to ensure that the acquisition of resources does not infringe on the rights of others to acquire resources, but is a weak restriction that makes it difficult to justify the acquisition of resources in practice. There are two mechanisms by which the proviso as it pertains to Nozick's entitlement theory breaks down.

Firstly, the proviso is a weak and vague restriction. It was written in an era when it seemed plausible that individuals would frequently be staking claim over new posessions in the wilderness, in particular parcels of land. However, in the modern setting, there are few unclaimed natural resources, and those that exist come under heavy contention for acquisition. The proviso does not provide a clear mechanism for dividing up the resources in this case, and it seems entirely unlikely that one can satisfy both aspects of the proviso concurrently.

Secondly, the proviso has a problem dealing with the issue of surplus value. According to the proviso, when an individual acquires a resource, they acquire it by instilling some valuable portion of themself into the resource. There is thus a fixed amount of value transferred onto the resource through the person's labor. However, in a free market like the one Nozick describes in his theory of entitlement justice, the value of a resource is not fixed, it is dictated by market forces. If an individual acquires a resource and then the value of that resource increases due to scarcity or high demand, then the individual can trade their resource and gain entitlement over property with a value greater than that which they instilled into their original acquisition.

These issues provide a strong challenge to Nozick's entitlement theory as they can result in disastrous consequences. Someone who is starving may "voluntarily" agree to trade property for food whose value is far below the value of the property, and per Nozick, this trade might be considered just. Critically, this does not spell the end for the entitlement theory of justice, but it does suggest that the underlying theory of property rights for a successful entitlement theory of justice must be more nuanced than the one Nozick proposed.

3.3 Instrumental Property Rights

More modern theories of entitlement have sought to address these issues by replacing the Lockean proviso with an alternate theory of property rights. Mack (1990) suggested that a theory of property rights derived from self-ownership would always be insufficient to justify a system of entitlement, proposing instead that property rights should be their own sepatate entity. van der Veen and

Van Parijs (1985) showed that entitlement systems existed on a spectrum such that the theory of property rights used could be tailored to the resource being distributed.

Regardless of the thoery of property rights used, the key to a successful entitlement theory is to ensure it can be justified using a consequentialist framework. Nozick's theory was a deontological theory, and as such it was heavily criticized for its potential to justify unacceptable outcomes as discussed above. Sen (1988) shows that while the interpretation of property rights as inerently valuable and inalienable leads to severe issues of poverty and hunger, the interpretation of property rights as *instrumental* rights, which are valuable only insofar as they lead to good outcomes, can be used to develop systems of entitlement without such issues. Instrumental property rights cannot supersede the demands of basic necessity for all agents, and so can be used to develop systems of entitlement which protect property rights while avoiding the issues of the Lockean proviso.

Combining these lessons, we realize that a successful modern theory of entitlement justice is one situated atop a theory of domain specific and instrumental property rights. For a given type of holding or resource, the theory of property rights must be tailored to the resource, and must be created and enforced with the full scope of its consequences in mind. "Re-situating" the Nozickean entitlement theory atop such a theory of property rights will allow us to develop a theory of entitlement justice which can be used to determine the justice of a distribution of resources in a modern society.

4 Entitlement Fairness

5 Conclusion

Aknowledgements

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2025-01-12.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. In *Fairness and Machine Learning*. fairmlbook.org, 2017. Online book, available at https://fairmlbook.org.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 2018. URL https://proceedings.mlr.press/v81/binns18a.html.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness, 2023. URL https://arxiv.org/abs/1808.00023.

Crime and Justice Research Alliance. Black men have higher rates of recidivism despite lower risk factors, 2023. URL https://crimeandjusticeresearchalliance.org/black-men-have-higher-rates-of-recidivism-despite-lower-risk-factors/. Accessed: 2025-01-21.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226. ACM, 2012.

Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the Machine Learning: The Debates Workshop at the 35th International Conference on Machine Learning (ICML)*, 2018. URL https://scholar.harvard.edu/files/bgreen/files/18-icmldebates.pdf.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323. Curran Associates, Inc., 2016.

M. C. Henberg. Nozick and rawls on historical versus end-state distribution. *The Southwestern Journal of Philosophy*, 8(2):77–84, 1977. ISSN 0038481X, 21541043. URL http://www.jstor.org/stable/43155157.

Corinna Hertweck, Christoph Heitz, and Michele Loi. What's distributive justice got to do with it? rethinking algorithmic fairness from the perspective of approximate justice, 2024. URL https://arxiv.org/abs/2407.12488.

Lily Hu. Normative facts and causal structure. *The Journal of Philosophy*, forthcoming. To appear.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there?, 2021. URL https://arxiv.org/abs/2105.01441.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018. URL https://arxiv.org/abs/1703.06856.

Eric Mack. Self-ownership and the right of property. *The Monist*, 73(4):519–543, 1990. doi: 10.5840/monist19907343.

Robert Nozick. *Anarchy, State, and Utopia*. Basic Books, New York, 1974. ISBN 978-0465097203. Proposes the entitlement theory of justice as a response to distributive justice theories like Rawls's.

Louis Pojman. Equality and desert. *Philosophy*, 72(282):549–570, 1997. ISSN 00318191, 1469817X. URL http://www.jstor.org/stable/3752010.

John Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, MA, revised edition edition, 1971. Original edition published in 1971, revised edition in 1999.

Amartya Sen. Equality of what? In Sterling M. McMurrin, editor, *The Tanner Lectures on Human Values*. University of Utah Press, Salt Lake City, Utah, 1979. Delivered at Stanford University, May 22, 1979. Available at: https://tannerlectures.utah.edu/_documents/a-to-z/s/sen80.pdf.

Amartya Sen. Property and hunger. *Economics and Philosophy*, 4, 1988.

Aaron Smith. Public attitudes toward computer algorithms, November 2018. URL https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. Accessed: 2025-01-12.

Robert J. van der Veen and Philippe Van Parijs. Entitlement theories of justice: From nozick to roemer and beyond. *Economics and Philosophy*, 1(1):69–81, 1985. doi: 10.1017/S0266267100001899.