Entitlement Justice and Measures of Algorithmic Fairness

Edward Speer
California Institute of Technology
espeer@caltech.edu

Llama-2-7b
Meta Platforms, Inc.

February, 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). This decision rule will determine whether or not agent a_n is allocated or denied some

good.

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

In the previous sections, we have shown the difficulty of connecting the philosophy of distributive justice to the practice of algorithmic fairness. Despite assertions in the literature that specific fairness measures relate to egalitarianism, we have shown that connections don't hold up under scrutiny. This difficulty warrants a reevaluation of the philosophical foundations of algorithmic fairness, and we propose that entitlement justice is a more apt framework for understanding the goals of fairness in algorithmic decision-making systems.

On an entitlement fairness account, an algorithmic decision-maker become a property rights oracle. The system at large is responsible for determining which individuals are entitled to a particular property. This requires two steps: first, we must know all of the factors which are morally relevant to property rights in the domain in question and be able to asign values to those factors for each individual in a population. Some of these factors we may be able to simple observe or measure directly, while many of these factors will have to be predicted algorithmically. For example, in the case of criminal parole, likelihood of recidivism is clearly relevant to one's entitlement to parole, but is something which must be predicted or learned from data. Second, we must implement an entitlement rule which determines whether or not an individual is entitled to a property based on the values of the morally relevant factors that we've predicted. Note that our decision-maker has now been split into two explicit components with very specific purposes. Following reasoning presented in Kuppler et al. (2021), refer to these steps as the prediction step and the decision step, respectively. The role of each component in the decision-making process now clearly implies the metric which should be used to evaluate the decision-maker.

The first component is a set of predictors which need to predict certain features of individuals

in a population. In the case of these predictors we are only interested in having as little prediction error as possible. We need the predictors to reflect the true properties of the individual in question so that we can use the factual information about them to make a decision about their property rights. Of course, a predictor will always be noisy, meaning there will be error in the prediction of the individual's properties. If this error is not distributed equally across the population, then we will consistently learn false information about a particular group which will lead to systemic errors in predicting property rights across groups. Therefore, each predictor must be evaluated using equalized odds fairness.

In step two, we implement an entitlement rule which determines whether or not an individual is entitled to a property based on the values of the morally relevant factors that we've predicted. We find ourselves in the fortunate condition that we have identified and predicted morally justified features to be relevant to the property rights in question. Therefore, our measurement of the decision step simply needs to ensure that *only* those features which are morally relevant to the property rights in question and were predicted in the prediction step are used in the decision step. In other words, given the set of morally relevant features, we need to ensure that the decision step is independent of other features which should not factor into the property rights decision. Therefore we must evaluate the decision step using counterfactual fairness.

In summary, an algorithmic decision maker implements entitlement fairness if it does the following:

- Given the set of covariates X, the decision maker algorithmically predicts the values of a morally justified set of features F for each individual in a population.
- Each predictor used in the prediction step satisfies equalized odds fairness.
- Based solely on the information in F, the decision step implements an entitlement rule which determines whether or not an individual is entitled to a property.
- The decision step satisfies a form of counterfactual fairness under which the decision step is independent of features not in F.

In the following sections, we will dive further into the connection of this framework to the philosophy of entitlement fairness and show the benefits of adopting this framework for algorithmic fairness.

4.1 End-state vs. Historical Fairness

Consider Nozick's distinction between end-state and historical theories of justice. In an end-state theory, to know if any allocation of one good to a person is just, we must know the broader distribution of all similar goods across society as a whole. This poses a major challenge to algorithmic fairness because it requires the decision-maker to have evolving data about individuals' stocks across society. This type of dataset is impossible to maintain and provide to the decision maker in practice. In contrast, historical theories of justice need only focus on the single particular good which is subject to a decision or allocation. While it could still be difficult to determine the full historical context of a good, it is at least a goal which is much more in line with the capabilities of an algorithmic system as we will show.

In the entitlement fairness framework, the decision-maker is a property rights oracle. The decision-maker is responsible for determining whether or not an individual is entitled to a particular property. This is a historical theory of justice because one's entitlement to a property is determined by some transactional history of the property. In the case of physical goods, this sort of structure is obvious — a good was acquired and then passed along through transfers of ownership into the hands of the current owner. Algorithmic decision makers seldom deal with physical goods, however. Instead, they tend to deal with opportunities. For example, college decisions allocate the opportunity to attend a particular university. Loan decisions allocate the opportunity purchase a house, start a business, or undertake other actions which require capital on credit. What sort of transactional history is there for these opportunities? And how does it impact one's entitlement to them?

In each of these cases, the transactional history is the history of opportunity given to the individual in question by society throughout their life. Consider some example cases:

- **College Admissions** (Access to education): Equality of opportunity in education is an entitlement justly acquired by each individual at birth. Since each individual is entitled to equal opportunity, there are no meaningful transfers of this opportunity between individuals. Therefore, the relevant history is only the acquisition combined with any infringements on the equality of opportunity. Under an entitlement framework, an individual who has been deprived of equal opportunity is entitled to rectification. Now note that we are able to predict based on, say, zip code and race, the extent to which an individual has been systemically deprived of equal opportunity. This means that we can predict the relevant history in the prediction step, and then use that information in our entitlement rule, successfully implementing a historical theory of justice.

- **Loan Decisions** (Access to capital): Unlike education, access to capital is not an automatic entitlement granted to individuals equally at birth. Instead, access to capital is based on merit. However, we will still employ a similar structure. An individual reaches a particular level of merit, acquiring the right to access capital. This merit is acquired through a history of work, education, and other actions and features relevant to the domain. However, past decisions about one's access to capital are certainly influenced by social categories and other factors which are not relevant to merit. Via some combination of income, race, and other factors, we can certainly predict whether one is more likely to have been unjustly deprived of access to capital, or unjustly rewarded with undeserved access to capital. This means that we can predict the relevant history in the prediction step, and then use that information in our entitlement rule to implement a historical theory of justice.

In each of these cases, it is not trivial to implement the relevant history, and requires extensive moral reasoning and domain-specific knowledge. However, the upshot is that using the entitlement fairness framework, we are able to implement a historical theory of justice in algorithmic decision-making systems, while implementing a truly end-state theory of justice would be impossible due to the required knowledge of the full distribution of goods across individuals in society.

4.2 Domain Specificity

In addition to the distinction between end-state and historical theories of justice, we also gain a new perspective on dmain specificity of fairness. In the algorithmic fairness literature, there are often broad attempts to reduce a multitude of fairness concerns to a single mathematical measure. However, this approach can reduce the visibility of the specific justice concerns of a particular domain, and possibly obscure the goals of fairness in that domain. For example, consider the COMPAS algorithm for assigning risk scores to individuals in the criminal justice system. The risk scores output by COMPAS are meant to represent likelihood of recidivism to inform decisions about bail, sentencing, and parole. A now infamous ProPublica article Angwin et al. (2016) showed that the COMPAS algorithm was biased against African Americans by measuring the false positive rate of the algorithm across racial groups. The false positive rate is clearly relevant to justice in this context, given that a false positive can lead to more frequent and longer incarcertations. However, Northepoint (the company which produces COMPAS) argued that their algorithm was fair because it had equal predictive accuracy across racial groups Anthony W. Flores (2016). Notice that on the entitlement fairness account, we would have detected the issues here — evaluation of the recidivism risk predictor would have revealed that it was not fair under equalized odds, meaning a different predictor would have been required to implement the decision rule. (Note that it is also highly unlikely that distributing sentences according to recidivism risk alone would qualify as a just rule of entitlement as well.)

Domain specificity is also critical to the selection of A, the set of protected attributes which should be considered in a fairness measure. In the algorithmic fairness literature, there is often a focus on selecting A to represent social categories, however there may be times when a social category is morally relevant to a particular decision. In the case of land ownership, for example, the ancestral heritage of an individual may be morally relevant to decisions about land allocation. In this case, it would be inappropriate to remove the attribute of ancestral heritage from the decision-making process. This effect is captured by the entitlement fairness framework. In deriving the set of features which must be predicted to implement the entitlement rule, we are forced to consider the domain-specific moral reasoning which is relevant to the property right in question. We then require that *only* those features not included in this set are excluded from the decision-making process. So for the case of land ownership, we may include a predictor of ancestal heritage in the prediction step, while still excluding race from the decision step — i.e, the only impact of race on the decision is through its impact on ancestral heritage.

Entitlement justice provides a framework for understanding the domain-specific considerations of fairness — as we discussed above, property rights are specific to the type of holding in question. Designing and building a system which respects property rights requires a deep understanding of the domain in which the system will be used. Understanding the fairness of a system in terms of entitlement justice therefore bakes in the domain-specific considerations of fairness into the design of the system.

4.3 A Framework for Entitlement Fairness

The presentation of entitlement fairness above is a high level description of how a system may achieve fairness. In this section, we present a framework for implementing entitlement fairness

in algorithmic decision-making systems, so that the goals of fairness can be achieved in practice.

1. **Define the property being allocated by the decision-making system**. This may be a physical good, an opportunity, a right, or some other type of property. There may be multiple framings of property when studying a particular problem domain. For example, in the home loan case, you may think that the property being allocated is access to capital, or you may think that the property being allocated is the opportunity to purchase a home. Both options will carry with them different moral reasoning processes, and it is important to consider one's options carefully to select the property with the most significant moral concerns.

2. **Define the set of attributes which are morally relevant to the entitlement to the property**. The implementor of the decision-making system must identify and *morally justify* the features of an individual which are relevant to their entitlement to the property in question. This set of attributes will vary widely depending on the property right in question.

3. **Implement fair predictors for the morally relevant features**. For each of the morally relevant features identified in the previous step, implement a predictor which predicts the value of that feature for each individual in the population. Each predictor must satisfy equalized odds fairness.

4. **Implement a fair entitlement rule**. Use the features predicted in step 3 to implement an entitlement rule which determines whether or not an individual is entitled to the property in question. There are many options for implementing an entitlement rule. One may implement something as straightforward as a decision tree or linear model based on the predicted features, or one may implement a less transparent model such as a neural network which learns the entitlement rule from some data. In any case, the entitlement rule must satisfy counterfactual fairness as described above.

5. **Publish moral reasoning and aknowledge limitations**. The final step in the framework is to publish the moral reasoning behind the morally relevant features and the entitlement rule. This will allow for public scrutiny of the rule and the identification of any limitations in the system.

4.4 Instrumental Algorithmic Decisions

Recall that under the entitlement fairness framework, the decision-maker is a property rights oracle. This means that regardless of the implementation or the problem domain, the outcome of the system describes whether or not an individual is entitled to a particular property under a system of entitlement. Returning to our discussion of instrumental property rights from section 3, we now observe a natural limitation of algorithmic decision-makers under the entitlement justice framework. Since, as Sen identified, property rights must be instrumental, so too must the output of the decision-maker be instrumental. This means that the decision of algorithmic decision-makers must not be absolute. There must be some flexibility in the decision-making

process to allow for the intervention on decisions based on higher priority issues in the broader sociotechnical system in which the decision-maker is embedded.

As an example, imagine that an algorithmic decision-maker is used to allocate college admissions decisions to students. The decision-maker determines that a particular student is entitled to admissions based on their academic merit in combination with, say, excellent essays and recommendations. The student has gained a property right to admission. However, say the university becomes aware of a higher priority issue — the student in question has serious disciplinary issues that they feel would place other students at their university in danger. The other students' rights to safety are more important than the student's property right, and therefore the university must be able to intervene and deny the student admission.

This type of intervention on the outputs of algorithmic decision-makers is often advocated for in the literature on socio-technical systems, and is considered to be a common sense measure, but critically, by considering the entitlement approach to algorithmic fairness, we see that this intervention emerges as a moral imperative, not just a pragmatic one. This is a significant benefit of the entitlement justice interpretation of algorithmic 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.

Cristopher T. Lowenkamp Anthony W. Flores, Kristin Bechtel. False positive, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Federal Probation*, 80(2):38–46, 2016.

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.