Entitlement Justice and Measures of Algorithmic Fairness

Edward Speer
California Institute of Technology
espeer@caltech.edu

Llama-2-7b
Meta Platforms, Inc.

March, 2025

1 Introduction

The rise of algorithmic decision making in the public sector has caused significant public concern. Algorithms increasingly make decisions that affect individuals' lives, from determining creditworthiness to predicting criminal recidivism, and the public has grown cautious 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 documented in the famed COMPAS case, these fears are not unfounded Angwin et al. (2016).

In response to these concerns, a growing body of research has focused on developing algorithmic fairness measures to evaluate and mitigate the biases in algorithmic decision making. A large number of different measures have been proposed Corbett-Davies et al. (2023) and applied to a wide range of problems. However, many questions remain unanswered about the theoretical foundations of these measures and their relationship to broader sociotechnical systems. In particular, the relationship of these measures to philosophically rigorous definitions of justice is not well understood.

Most recently, Binns (2018), Hertweck et al. (2024), and Kuppler et al. (2021) have explored this relationship, attempting to make sense of the theoretical foundations of algorithmic fairness measures through the lens of distributive theories of justice in political philosophy. A distributive theory of justice is a normative framework that provides principles and criteria for allocating resources, benefits, and burdens among individuals or groups within a society, with the aim of achieving a just and fair distribution. At a first glance, it is appealing to compare such a scheme to the problem of algorithmic fairness — encode the constraints on the allocation into the fairness measure, and optimize for the measure to achieve a fair allocation. However, this approach demands further investigation. How do we encode the constraints of a distributive theory of justice into a fairness measure? How well do the measures in the literature capture competing accounts of distributive justice?

The papers mentioned above all argue that existing fairness metrics implement different forms of egalitarian justice, primarily luck egalitarianism. In this paper, we will carefully examine the relationship between algorithmic fairness and distributive justice, and argue that the existing literature is too narrow its scope. We will argue that algorithmic fairness should be understood

through the lens of entitlement justice, a distributive theory of justice that has been largely overlooked in the literature. We will argue that entitlement theory, under which justice is rooted in the idea of respecting individuals' property rights, offers a more nuanced and context-sensitive understanding of algorithmic fariness. We argue that by conceptualizing the algorithmic fairness problem through the lens of entitlement justice, system designers are forced to clearly present the inherent normative reasoning and values endorsed by their systems.

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 there is a decision problem under scrutiny that can be formulated using the following formalism. There is a population of individual $I = \{i_1, i_2, \ldots, i_n\}$ over whom we must distribute a resource R. Following Kuppler et al. (2021), we define a *decision rule* as a mapping $d : I \rightarrow \{0, 1\}$ which determines whether or not a given individual i_n receives an amount of resource R. An algorithmic deicion-maker on this account is a decision rule d which is implemented by an algorithm.

At this time, it worth defining what we mean by a resource R more carefully. For this purpose, we will use Rawls' definition of a primary good Rawls (1971). For Rawls, these are a specific category of goods which consist of

1. Basic rights and liberties
2. Freedom of movement and free choice among a wide range of occupations
3. Powers of office and positions of responsibility
4. Income and wealth

For our purposes, any such primary good is fit to be considered a resource in our decision problem. The typical problem domains for algorithmic fairness fit neatly within these categories. For example, college admissions algorithms adjudicate freedom of occupation, bank loan algorithms distribute wealth, promotion algorithms in the worksplace divy up positions of responsibility, and prison parole decisions determine freedom of movement.

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 typical presentation of algorithmic fairness, we are given a population of individuals $I = \{i_1, i_2, \ldots, i_n\}$ who have a set of covariates P. There are a set of morally protected covariates, $M \subseteq P$ of an individual, an observed set of covariates $X \subseteq P$ for each individual, and therefore a set of observed morally protected covariates $A = X \cap M$.

Our algorithm is a classification function $f : X \rightarrow 0, 1$ mapping covariates of each individual to a binary outcome Corbett-Davies et al. (2023). A fairness metric is a function over f and A designed to measure the extent to which f obeys some constraint on the distribution of outcomes with respect to A.

With respect to the decision problem described above, we now have a model for the full decision-making process. For each individual competing for a resource, we have a decision rule which says to measure a set of covariates X for each individual, and then apply f to X to determine whether or not to allocate R to that individual.

Note the strict difference between the decision rule d and the function f. The decision rule operates on the entire individual i_n, while the classification function operates only on the observed covariates from the individual. This distinction is important because it demonstrates a key limitation of algorithmic fairness measures. One might be concerned that a decision rule is unfair despite a fair classification function if the covariates input into the classifcation function are not a complete or accurate representation of the individual.

As a running example, consider the case of a hiring algorithm. The decision problem is to map a set of applicants to hiring decision. Our algorithm must implement a function to do so based only on the personal information presented in their resumes, which is their education, experience, race, and gender. We want to ensure that our hiring practices are fair with respect to race, and gender. This gives us the following setup:

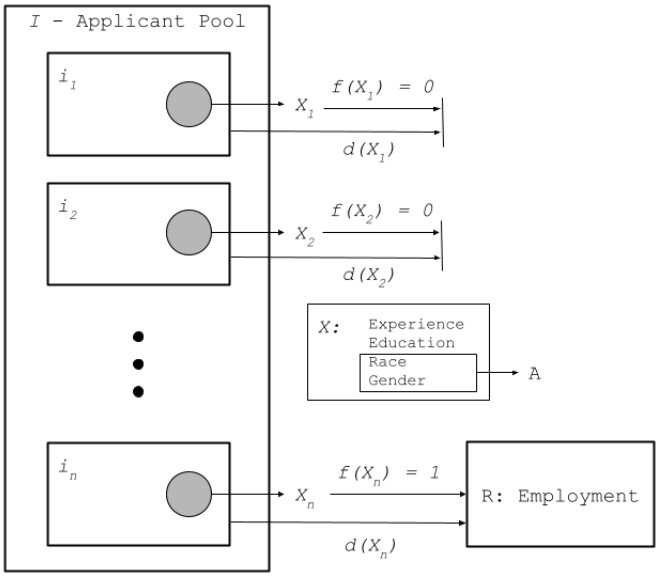

Fig. 1: A hiring decision algorithm as a decision problem

where a fairness measure will operate exclusively on the mapping from experience, education, race, and gender to hiring decisions. Whether or not these factors are a good reputation of the applicants is left unaddressed by the fairness measure.

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

Definition 1. *Fairness Through Unawarenss — f satisfies fairness through unawareness iff*

$$A = \emptyset$$

In other words, the classification function may not receive any morally protected covariates as inputs. So, for our hiring admissions case, we would be forced to remove race and gender from X and only consider education and experience.

This notion is intuitively appealing — how can the algorithm dicriminate against me based on my race if it doesn't know my race? However, it is clear that this measure is not sufficient to ensure fairness. For example, if I attended a historically black college, my education may function as a proxy for my race, allowing bias to remain in the algorithm.

Definition 2. *Demographic Parity — f satisfies demographic parity iff*

$$P[f(X) = 1|A = a] = P[f(X) = 1] \ \forall a \in A$$

Dwork et al. (2012).

Demographic parity holds that the probability of a positive output ($f(X) = 1$) should be statistically independent of the protected attributes. This is an easily understandable and measurable criteria for fairness. At a first look, it is appealing — the same number of individuals from each

race will be successful in seeking jobs at a particular company. However, a close look reveals difficulties.

Under demographic parity, we must balance the probability of success between groups, which becomes very difficult when the base rates of success are unequal. For example, if women are much more qualified for a job on average than men, then demographic parity will require that we hire less qualified men in place of more qualified women in order to balance the probability of success between gender groups. In other words, my false positive rate will be very high for men while my false negative rate will be very high for women, creating a severely unfair practice Barocas et al. (2017).

Definition 3. *Equalized Odds — f satisfies equalized odds if given true outcomes D over all individuals, we have*

$$P[Y = 1|A = a, D = 1] = P[Y = 1|A = b, D = 1] \ \forall a, b \in A$$

$$P[Y = 1|A = a, D = 0] = P[Y = 1|A = b, D = 0] \ \forall a, b \in A$$

Hardt et al. (2016).

Equalized odds requires that the true positive and false positive rates be balanced between groups. This is often thought to ensure there is no disparate mistreatment across groups. In our hirign case, for example, equalized odds ensures that no one racial group more likely to be falsely rejected or erroneously hired than another. If I am rejected from a position despite being qualified, I can be assured that had my race been different, the probability of this outcome would have been the same.

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, one could achieve equal false positive rates by *adding* false positives to the white portion of the dataset, resulting in an increase in the number of white prisoners receiving parole undeservingly. This was exactly the case in the COMPAS algorithm Angwin et al. (2016). COMPAS was calibrated to have equal predictive accuracy across racial groups, but this resulted in a higher false positive rate for black defendants and a much lower false positive rate for white defendants due to unequal base rates.

Definition 4. *Counterfactual Fairness — f satisfies counterfactual fairness iff*

$$P[f(X) = 1|do(A = a)] = P[f(X) = 1|do(A = b)] \ \forall a, b \in A$$

Kusner et al. (2018). Where the do operator denotes an intervention on the variable A.

Borrowing from the language of causal inference, counterfactual fairness posits that the protected attributes may not have any causal effect on the outcome of te classification function.

This is a highly appealing notion of fairness. If my protected characteristics do not in any way cause the outcome of the decision rule, then it is difficult to argue that I have been discriminated against. However, this measure is difficult to implement in practice.

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 critiques reveal that there is no one-size-fits-all solution to the problem of algorithmic fairness. Each measure has its own strengths and weaknesses, and the choice of measure will depend on the specific context in which the algorithm is being applied. However, how should one select a measure? What are the normative considerations that should guide this choice? In hopes of developing a more nuanced and structured approach to these questions, we turn to work in the philosophy of distributive justice.

2.2 Theories of Distributive Justice

Within the formalism we have presented, the role of a theory of distributive justice is to define what constitutes a fair distribution rule. While algorithmic fairness measures operate on the classification function which is one component of the decision rule, distributive justice operates at one level higher, defining the rule decision rule in which the classification function is embedded.

Decision rules posited by theories of distributive justice take on the following general form: Allocate R to i_n iff i_n has property y. Several conventional theories of distributive justice have been proposed in the literature, and we will discuss those which have been previously proposed to be implemented by algorithmic fairness measures here. For a more exhaustive list of theories available, see Kuppler et al. (2021)

Definition 5. *Liberal Egalitarianism — liberal egalitarianism posits that equal stocks of resources should be maintained across society Rawls (1971).*

Equally restated in the formalism we have presented, egalitarianism posits that R should be allocated to i_n iff doing so minimizes the overall inequality across stocks in the population. Thus in this case, y is the property of *lacking* R relative to the population. This approach is appealing in its simplicity and clarity, and in its ability to ensure that all individuals have access to the same resources.

Note that the precise currency of liberal 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 6. *Luck Egalitarianism — Luck egalitarianism posits that individuals should receive resources in proportion to their choices and actions, but equally with respect to their features which are outside of their control Knight (2013).*

Luck egalitarianism posits that R should be allocated to i_n iff i_n deserves R based on some set of their choices and actions which do not include features outside of their control. This approach is appealing on the grounds of personal responsibility and autonomy. If I make choices which result in my own deprivation of R, then I am responsible for that deprivation. However, I cannot be deprived of R on the grounds of features I do not control, such as my race, gender, or talent.

Luck egalitarianism is often critiqued for being highly subjective and difficult to measure. What constitutes a choice or action is not always clear, and the line between what is and is not within an individual's control is often blurry. Establishing responsibility is already a difficult task in legal philosophy, and operationalizing this theory in practice is likely to be fraught with difficulties.

Definition 7. *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 R which is sufficient to meet some threshold level of well-being. Thus y is defined as the property of *needing* R. This sort of approach protects the most vulnerable members of society and ensures that all individuals have access to the basic goods they need to survive.

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. It therefore remains unclear how to operationalize this theory in practice in many cases.

Definition 8. *Desert — Moral desert is the idea that individuals should receive resources in proportion to their moral merit Pojman (1997).*

Theories of desert therefore set y to be some operationalization of moral merit. Under the right definition of moral merit, this approach shows promise — it is intuitive that someone who has worked hard and done good works across society deserves reimbursement or insurance of positive outcomes. However, 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.

The main differentiating factor between these theories is the property y which they posit as the basis for the decision rule. In seeking to understand algorithmic fairness through the lens of theories of distributive justice, we must therefore ask how, if at all, the fairness measures we have presented place constraints on the property y of the decision rule. Especially given that the fairness measures operate on the cliassification function which is a component of the decison rule, to what extent are they capable of enforcing constraints over the full decision rule? So far as they do enforce constraints on y, how well do they capture the constraints actually posited by the common theories of distributive justice enumerated here?

2.3 Relating Algorithmic Fairness to Distributive Justice

We have now developed the key components of our formalism needed to connect algorithmic fairness to theories of distributive justice. In the algorithmic decision-making process, we have a

decision rule which is composed of the observation of a set of covariates X for each individual, the application of a classification function to X to determine whether or not to allocate a resource to that individual, and the allocation of the resource to an individual if it is dictated by the classification function. Theories of distributive justice define what a fair decision rule is, and in doing so, they define the property y which determines whether or not an individual should receive the resource. Algorithmic fairness measures, on the other hand, operate only on the classification function. Therefore, it doesn't make sense to argue that an algorithmic fairness measure directly implements a theory of distributive justice. Instead, we must evaluate the entire decision rule, including the mapping from each individual to their observed covariates to determine whether or not the correct constraints are being placed on y for a given decision rule.

As a concrete example, return to the case of a hiring algorithm. If X is modified to exclude education and experience, and to include, say, enjoyment of Mexican food, then regardless of what fairness measure is satisfied by f, d will clearly be unjust according to any theory of distributive justice. This example demonstrates that the constraints placed on f by a fairness measure are not sufficient to ensure that the full decision rule d is complies with a theory of distributive justice, but must operate in conjunction with the constraints placed on X.

Once granted an appropriate X as input to f, the role of f becomes to predict the y desired by the chosen theory of distributive justice. The role of the fairness measure is then to measure the extent to which f is successful in doing so. For example, if we are operating under a theory of desert, then f should predict the moral merit of each individual, and the fairness measure should measure the extent to which f is successful in doing so. If f is successful in predicting moral merit, then the decision rule d will be fair according to the theory of desert. Let us then briefly examine the fairness measures we have discussed in this light, to see what constraints they place on y given an appropriate set of observed covariates as input.

- Fairness Through Unawareness: Fairness through unawareness places no constraints on y. It only requires that the classification function not receive any morally protected covariates as input. Since morally protected covariates can be derived from those covariates which are not protected, this measure is essnetially vacuous with respect to the full decision rule.

- Demographic Parity: Demographic parity mandates that y should not deppend on the morally protected covariates. This is too weak a constraint to fully implement any theory of distributive justice. For example, this condition is *necessary* but not *sufficient* for liberal egalitarianism, luck egalitarianism, and sufficientarianism, but is not strong enough to fully enforce any of these theories.

- Equalized Odds: Equalized odds dictates that y should not depend on the morally protected covariates given the true outcomes. In other words, y is a property of deservingness which is independent of the morally protected covariates. This *could* implement moral dessert under a particular definition of moral merit in which the morally protected covariates are excluded from the operationalization of moral merit and the true outcomes are the direct result of moral merit. However, this is undesireably working backwards from the fairness measure to the theory of distributive justice.

- Counterfactual Fairness: Counterfactual fairness requires that y is a property of deservingness which is causally independent of the morally protected covariates. This measure again implements moral dessert under a hyper-specific operationalization that is unproductive for our purposes.

This analysis reveals that the fairness measures we have discussed are not sufficient to enforce theories of distributive justice. They place constraints on the classification function which are too weak to fully enforce the constraints placed on the full decision rule by theories of distributive justice, except in hyper-specific cases which are unproductive for justifying the use of the fairness measure. What's missing in each case is *detail*: How does an algorithmic fairness measure enforce the normative contraints about how y is derived from X? This suggests a new account is needed to synthesize these two fields and to provide a more nuanced and structured approach to connecting algorithmic fairness to theories of distributive justice. In the following sections, we will develop such an account through the lens entitlement justice, which has been largely overlooked in the literature on algorithmic fairness.

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 Justification of Acquisition

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.

Other accounts of entitlement justice have used different mechanisms to justify the acquisition of resources.. Mack (1990) proposed that the acquisition of resources could be justified as a separate unalienable right that all individuals possess. van der Veen and Van Parijs (1985) proposed that the acquisition of resources could be justified consequentially by the net utility that the acquisition

brings to society In general, Van Der Veen showed that given a particular type of holding, one can specify a theory of entitlement justice with a corresponding utilitarian theory of acquisition that can be used as a basis for determining entitlements.

3.3 Critiques of Entitlement

Nozick's entitlement theory is heavily criticized on its foundation in the Lockean proviso. The final clause of the proviso provides a restriction on the extent to which resources can be acquired, 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. Besides the proviso, Nozick's theory may be criticized for its potential to justify unacceptable outcomes through transfer. For example, 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 as well as the restrictions on the types of transfers it can justify must be more nuanced than what Nozick proposed.

3.4 Instrumental Property Rights

As described, ore modern theories of entitlement have sought to address these issues by replacing the Lockean proviso with an alternate theory of property rights. 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. For example, the theory of property rights used to distribute land could be different from the theory of property rights used to distribute food or water.

Regardless of the thoery of property rights used, to overcome the challenges of Nozick's theory, one must prevent an entitlement theory from justifying morally unacceptable outcomes. 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 types of 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 becomes 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 assign values to those factors for each individual in a population. Some of these factors we may be able to simply 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. We now have a pipeline which appears as follows:

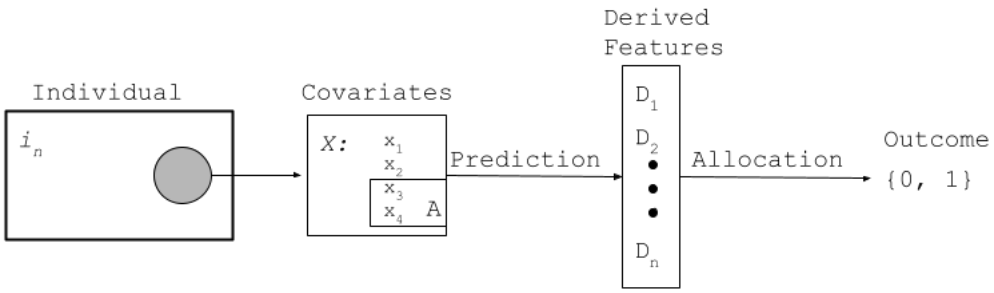

Fig. 2: Algorithmic Decision-Making Pipeline

In order to implement this pipeline, one must have a clear understanding of the property right in question, and the morally relevant factors which determine entitlement to that property. Note that these factors may be related to sensitive social categories. For example, if I am determining property rights to land, ancestral heritage may be a morally relevant factor. Then we know that protected attributes are allowed to play a role in the prediction step, so long as they are morally relevant to the property right in question and the moral reasoning behind their inclusion is made explicit. In the decision step, however, these attributes must be excluded, and the decision must be made based only on the predicted morally relevant factors, since only these are the features which have been shown to be relevant to the property right in question.

Therefore, measuring the fairness of an algorithmic decision-making system under the entitlement justice framework means that we need to require that the protected characteristics of an individual act on the decision *only* through those features which are morally relevant to the property right in question. If we can ensure this, then we have a system in which sensitive characteristics of an individual are used to assess their entitlement to a particular property in only ways which are explicitly morally justified, and otherwise do not play a role in the decision-making process. This notion is precisely formulated as the following specific form of counterfactual fairness:

Definition 9. *Given a set of covariates X, a set of protected attributes A, a set of morally justified derived features D, and a classification function f, a decision-making system satisfies* ***entitlement fairness*** *iff*

$$P[f(X) = 1|D = d, do(A = a)] = P[f(x) = 1|D = d, do(A = b)] \ \forall a, b \in A$$

In other words, the morally justified feature set D acts as a mediator between the protected attributes A and the decision $f(X)$, and no further influence of A on $f(x)$ is allowed. In the following discussion, we will show the advantages of adopting this perspective on 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 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. These predictors may be simple or complex, depending on the nature of the feature and the data available. The important thing is that the predictors must be able to predict the morally relevant features with some degree of accuracy.

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.

5. **Verify the fairness of the system**. Once the system has been implemented, it is important to verify that the system is fair. This may involve testing the system on a holdout dataset, or using a fairness measure to verify that the system is fair. The fairness measure should be based on the definition of entitlement fairness given above.

6. **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

In this paper, we have presented a novel approach to the problem of justifying algorithmic fairness criteria. We have presented a formalism for discussion of algorithmic fairness that is grounded in the theory of distributive justice, and used that formalism to demonstrate the limitations of existing fairness criteria with respect to distributive justice. In order to overcome these limitations, we have suggested turning to the theory of entitlement justice, which has been largely overlooked in this area of inquiry. Leveraging entitlement theory as a political philosophy, we have proposed a new fairness criterion, the *entitlement fairness criterion*, which has a clear and structured connection to theories of justice derived from political philosophy. The benefits of this approach are threefold.

First, the entitlement fairness criterion forced algorithm designers to confront the moral reasoning inherent in their design choices. In grounding fairness in entitlement justice, designers are made to present a clear account of property rights specific to the problem domain in question. This is a significant improvement over existing fairness criteria, which often rely on vague notions of fairness that are not grounded in any particular theory of justice, and may hide normative assumptions that are not immediately apparent.

Second, the entitlement fairness criterion provides a clear and structured framework for

evaluating the fairness of algorithms. Rather than choosing from an ad hoc list of fairness criteria, designers can use the entitlement fairness criterion to evaluate the fairness of their algorithms in a systematic and consistent manner. This allows for a more rigorous evaluation of the fairness of algorithms, and can help to identify and correct biases that may be present in the design of algorithms.

Finally, the entitlement fairness criterion provides a principled basis for human intervention in the decision-making process of algorithms. By grounding fairness in property rights which are instrumental in nature, the entitlement fairness criterion explicitly demands human oversight and correction in cases where the algorithm fails to respect rights which take higher precedence than property rights do. This bakes in a broader awareness and accountability for the full range of ethical considerations that may be at play in the design and implementation of sociotechnical systems.

In conclusion, we believe that the entitlement fairness criterion offers a promising new approach to the problem of justifying algorithmic fairness criteria. Future work in this area should focus on developing a more detailed account of how to formulate property rights in specific problem domains, and on exploring how the entitlement fairness criterion can be used to evaluate the fairness of algorithms in production. We hope that this paper will inspire further research in this area, and that the entitlement fairness criterion will become a valuable tool for ensuring that algorithms are designed and deployed in a fair and just manner.

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.

Carl Knight. Luck egalitarianism. *Philosophy Compass*, 8(10):924–934, 2013. doi: https://doi.org/10.1111/phc3.12077. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12077.

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.