A Systems Thinking Approach to Algorithmic Fairness

Chris Lam
Epistamai
Apex, NC, 27502
chris@epistam.ai
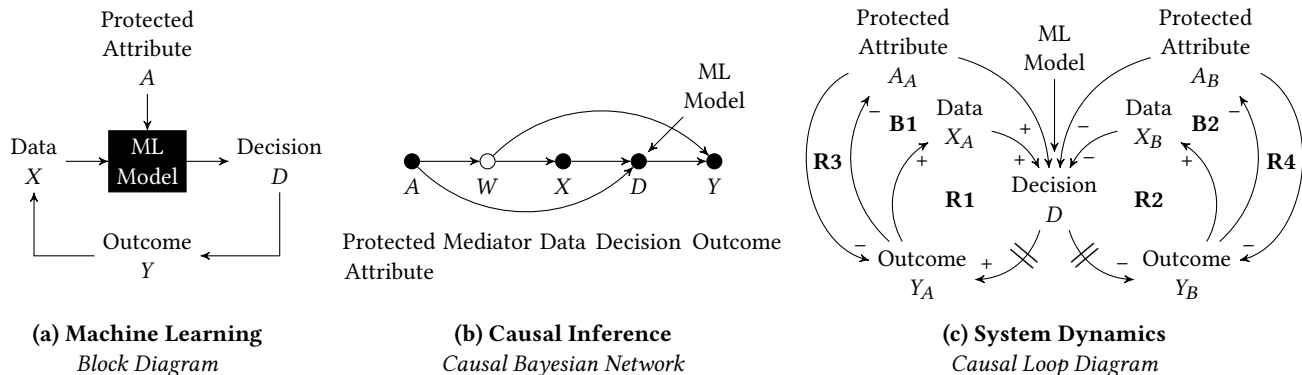
(a) Machine Learning
Block Diagram

(b) Causal Inference
Causal Bayesian Network

(c) System Dynamics
Causal Loop Diagram

Figure 1: Three representations of algorithmic fairness

ABSTRACT

Systems thinking provides us with a way to model the algorithmic fairness problem by allowing us to encode prior knowledge and assumptions about where we believe bias might exist in the data generating process. We can then encode these beliefs as a series of causal graphs, enabling us to link AI/ML systems to politics and the law. This allows us to combine techniques from machine learning, causal inference, and system dynamics in order to capture different emergent aspects of the fairness problem. We can use systems thinking to help policymakers on both sides of the political aisle to understand the complex trade-offs that exist from different types of fairness policies, providing a sociotechnical foundation for designing AI policy that is aligned to their political agendas and with society's shared democratic values.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; **Probabilistic reasoning**; **Reasoning about belief and knowledge**; • **Applied computing** → Law; Sociology; Economics; Psychology; • **Social and professional topics** → Governmental regulations.

KEYWORDS

systems thinking, machine learning, causal inference, system dynamics, emergence, bias, fairness, discrimination, disparate treatment, disparate impact

1 INTRODUCTION

For several years, AI researchers have struggled with how to build AI/ML systems that are fair and aligned to society's shared democratic values. Purely technical approaches that rely on correlation alone have fallen far short of what is needed to solve this difficult problem. Systems thinking offers a novel approach to the algorithmic fairness problem by allowing us to use causality to combine complex ideas from the social sciences and computer science into a comprehensive and integrated theory. By leveraging prior knowledge and assumptions about where we believe bias might exist in the data generating process, we can help policymakers on both sides of the aisle to design AI policy that is aligned to their political agendas.

To address the algorithmic fairness problem, we need to take a sociotechnical approach. The first part of this paper will briefly cover concepts from the social sciences to address the controversial question: Why do we believe disparities exist across protected groups like race or gender? This will give us insights into the political debate around fairness and why policymakers fundamentally disagree on how to close those disparities.

We will then turn our attention to the technical aspects of the fairness problem, which forms the bulk of this paper. We will use systems thinking to show how to combine analytical techniques from machine learning, causal inference, and system dynamics to model different facets of the fairness problem. Machine learning is necessary for building AI systems that can make fair predictions. Causal inference allows us to visualize different forms of bias, showing us what interventions are needed to make a machine learning model fair. System dynamics allows us to model bias from the data generating process, providing counterintuitive explanations as to why disparities continue to persist across protected groups. Finally, we will demonstrate how these three techniques can form the basis of a new causal hierarchy for modeling algorithmic fairness.

2 RELATED WORK

The field of algorithmic fairness is a relatively new field that has drawn significant attention since the release of a ProPublica article that claimed racial bias in COMPAS, a software tool that is widely

used in the US court system to predict recidivism[1]. Since then, there have been multiple attempts to align the legal definitions of discrimination (e.g. disparate treatment and disparate impact in the US or direct discrimination and indirect discrimination in the EU) with their equivalent mathematical definitions. But these approaches, which have oftentimes relied on statistical correlation, have not been successful. Narayanan for example identified nearly two dozen fairness metrics, none of which could fully align to the legal definitions of discrimination[30].

Pearl has argued for a few decades that the discrimination problem is fundamentally a causality problem[34]. He has argued that fairness can be modeled using causal Bayesian networks (CBNs) as a series of direct, indirect, and spurious effects. This work was later expanded upon by Plecko and Bareinboim, who proposed the idea of a standard fairness model[37]. They argued that there exists a causal template through which entire classes of fairness models can be derived using a small number of modeling assumptions. Kusner et al. argued for the use of counterfactuals in modeling fairness[24]. Chiappa and Isaac showed that certain paths within a CBN could be labeled as fair or unfair[9]. Further attempts at building causal fairness models were made by Nilforoshan et al.[31] and Kilbertus et al.[21]. Barocas et al. devoted a chapter in their textbook on fairness and machine learning to discuss how causality might be used to model fairness[3].

However, all of their approaches were limited due to their inability to construct a valid causal model of the problem. This is because their approaches did not fully take into account the social and political aspects of the fairness problem. Society has different ideological beliefs about what is fair and unfair based on different sets of assumptions about where bias is within the data generating process. Our paper addresses this issue by encoding society's different mental models of bias into causal models using first principles. By taking into account a diversity of perspectives, we can understand the normative aspects of the fairness problem. This is necessary for linking machine learning to politics and ultimately to the law.

Another key difference is that their causal approaches relied solely on causal inference, which has limitations in its ability to model the dynamics of the data generating process. Our approach leverages systems thinking and system dynamics to gain a fuller understanding of the data generating process through the modeling of feedback loops. Martin has also explored modeling AI bias using community-based system dynamics[15, 16]. Furthermore, we link supervised machine learning to causal inference by recognizing that a CBN can be used as a higher level abstraction of a black box ML model. By bringing supervised machine learning, causal inference, and systems thinking together, we can model different aspects of the algorithmic fairness problem.

3 TOWARDS A UNIFIED THEORY OF FAIRNESS

To build fair AI/ML systems, we begin from first principles with the strong assumption that all humans have free will[7]. This forms the basis of moral responsibility[46]. If we humans have free will, then we have the agency to make our own decisions. These decisions will lead to certain outcomes, and we may be held morally responsible for those outcomes.

Most of the time, we humans make decisions that only affect our own selves. We can be held personally responsible for the decisions that we make for ourselves. But sometimes, we humans make decisions that affect other people. This is especially true for humans who are in positions of great power. We may be held socially responsible for the decisions that we make for others.

Sometimes the decisions that we make will lead to good outcomes for other people. In these cases, they might assign us praise and reward. On the other hand, sometimes the decisions that we make will lead to bad outcomes for other people. In these cases, they might assign us blame and punishment. This concept is known as moral desert[17]. We humans have a natural tendency to accept praise and reward, and to reject blame and punishment.

But sociologists would note that we are never fully free to make our own decisions. This is because social structures can influence our agency[2]. Social structures define our relationships with the various institutions of society and may include our economic systems, legal frameworks, cultural norms, and political institutions. Agency defines our ability to act independently and to freely make our own decisions. Social structures place a limit on the the choices and opportunities that are available to us humans. Those who are economically disadvantaged will encounter greater limits to their agency due to social structures compared to those who are economically advantaged.

To better understand human behavior, we turn to a concept in psychology called the locus of control. This was coined by Rotter and refers to the degree of belief that people have control over the outcomes of events in their own lives[38]. Humans with an internal locus of control believe that they are in control of their own lives, while those with an external locus of control believe that their lives are controlled by outside factors. These concepts might also be related to whether someone adopts a growth mindset[11] or a victimhood mindset[18].

Psychologists provide another key concept for fairness called attribution theory, which is used by humans to interpret and assign causes for events. First proposed by Heider[14] and later expanded by Kelley[19] and Weiner[48], attribution theory examines whether attributions should be assigned to internal (dispositional) attributions or external (situational) attributions. The internal attributions may be due to behavior or personal traits, while the external attributions may be due to situational or environmental factors. This is essential for determining whether we believe an outcome is biased and therefore unfair.

For example, let's say that a student did poorly on an exam. An internal attribution might argue that the student was lazy and should have spent more time studying. In this case, the student is to blame for the poor results and thus the test was unbiased and fair. An external attribution might argue that the test was too hard or that the material was not taught well. In this case, the professor is to blame for the poor results and thus the test was biased and unfair.

To determine whether an outcome is biased and unfair, all we need to know is whether we would attribute a bad outcome as being internal or external. This makes algorithmic fairness a type of credit assignment problem. But since we usually focus on bad

outcomes, this oftentimes becomes a blame assignment problem. That being said, identifying attribution is much harder in practice, as we will show later when we model fairness using system dynamics. Both internal and external factors may influence and reinforce each other, leading to vicious feedback loops. This becomes a particularly difficult problem when trying to explain disparities across protected groups like race, which we describe in greater detail in Section 6.2.

This causality dilemma between attributing disparities to internal or external factors ultimately leads to the controversial and heated political debate over how best to reduce disparities across protected groups[10, 20, 26, 29, 43]. Those on the political left are more likely to attribute disparities as being due to external factors, which provides the moral justification for strong government interventions like affirmative action or DEI. The left also tends to focus more on legal forms of discrimination like disparate impact or indirect discrimination. Those on the political right are more likely to attribute disparities as being due to internal factors, which provides the moral justification for limited government intervention like fairness through unawareness. The right also tends to focus more on legal forms of discrimination like disparate treatment or direct discrimination. This ideological debate is fundamental to understanding algorithmic fairness, but has been largely missing from the academic literature.

We will use systems thinking to encode these different ideological beliefs into a series of causal graphs as shown in Figure 1. This will help us to understand what types of interventions policymakers on both sides of the political aisle might want to make to ensure that a machine learning model is fair. Finally, this could give us insights into how to build machine learning models that are compliant with antidiscrimination laws.

4 MODELING FAIRNESS AS A MACHINE LEARNING PROBLEM

Before we build a causal representation of the fairness problem, we will start by modeling fairness as a supervised machine learning problem using a simple block diagram as shown in Figure 1a. In the center is a black box machine learning model that receives data X as input and returns decision D as output. The decision D causes an outcome Y that could potentially be recorded as future data X, thus creating a feedback loop. The key question that we need to address is whether the protected attribute A should be used in a machine learning model and if so, how?

5 MODELING FAIRNESS AS A CAUSAL INFERENCE PROBLEM

We can transform the block diagram from Figure 1a into a causal Bayesian network (CBN) as shown in Figure 1b. CBNs were pioneered by Pearl[33, 35, 36] and allow us to symbolically encode knowledge and assumptions about causal effects between variables, which are shown in this graph as nodes. The first node of the CBN represents a protected attribute A, which has a direct effect on a mediator W ($A \rightarrow W$). The mediator W has a direct effect on an outcome Y ($W \rightarrow Y$). We assume that the mediator W completely explains away the causal effect between the protected attribute A and the outcome Y, resulting in conditional independence ($Y \perp\!\!\!\perp A|W$). However, the mediator W might be considered biased by some

people if they attribute the disparities across protected groups as being external. The mediator W is filled white because it is latent, but it can still be partially measured as data X ($W \rightarrow X$). Data X is then fed into a machine learning model that is embedded inside the node for decision D ($X \rightarrow D$). Finally, the decision D also has a direct effect on an outcome Y ($D \rightarrow Y$). CBNs are represented by directed acyclic graphs (DAGs), which means that we cannot use them to model feedback loops. Examples of what each of these variables could represent are shown in Table 1. For each of these applications, the protected attribute A could represent race or gender for example.

Figure 2 shows a series of CBNs modeling different forms of bias, fairness, and discrimination. Each CBN captures a different way of framing the fairness problem based on an individual's prior beliefs about where bias might exist in a model. On the left, we have four different forms of fairness, which we will later show can be mapped to different points on the political spectrum. On the right, we have four different forms of discrimination, which can be mapped directly to antidiscrimination law.

The first form of fairness that we will look at is **fairness through unawareness** (Figure 2c). This is the simplest form of fairness. We begin with the belief that the data is fundamentally unbiased (internal attribution), which is why all of the nodes in the graph are colored black. In this form of fairness, there is no edge between the protected attribute A and the decision D. Therefore, the model is blind to an individual's membership to a protected group A and therefore cannot treat any individual differently on that basis. We would say that the protected attribute A has no direct effect on the decision D, making the decision fair.

The advantage of fairness through unawareness is that it avoids **overt discrimination** (Figure 2b). If a model uses knowledge of an individual being a member of a disadvantaged group against them in its decision D, then it would clearly cause disparate treatment or direct discrimination. We show this as a form of bias where the edge from the protected attribute A to decision D is colored red ($A \rightarrow D$). This leads to a biased decision D, which in turn causes a biased outcome Y ($D \rightarrow Y$). Note that in these CBNs, bias is represented by colored nodes and flows across the entire graph. We would say that the protected attribute A has a negative direct effect on the decision D ($A \rightarrow D$), leading to harmful discrimination.

But if we instead believe that the data is fundamentally biased (external attribution), then fairness through unawareness could lead to **covert discrimination** (Figure 2d). In this scenario, there is bias between the protected attribute A and the mediator W ($A \rightarrow W$), which flows across the entire graph and causes all of the remaining nodes in the graph to be colored red. This bias can be attributed to structural or systemic forms of discrimination that unfairly disadvantages one group over another. For example, if a disadvantaged group was overtly discriminated against in the past, then the present data X may still reflect that negative bias. By removing the edge from the protected attribute A to the decision D, it could reinforce one group's prior success over another. Covert discrimination could form the basis of a disparate impact or indirect discrimination claim. We would say that the protected attribute A has a negative indirect effect on the decision D ($A \rightarrow W \rightarrow X \rightarrow D$), leading to harmful discrimination.

Table 1: Examples of high stakes applications for algorithmic fairness

Application	Mediator W	Data X	Decision D	Outcome Y
Credit	Creditworthiness	Income, credit history	Deny loan?	Loan default?
Insurance	Insurability	Demographics, claims history	Deny policy?	Claim filing?
Housing	Home value	Location, size, condition	Home valuation?	Final sales price?
Employment	Qualifications	Experience, education	Screen resume?	Turnover?
Education	Merit	Grades, test scores	Reject applicant?	Graduation?
Research	Merit	Quality, clarity, originality	Reject paper?	No. of citations?

(a) Fairness through supremacism

(b) Overt discrimination

(c) Fairness through unawareness

(d) Covert discrimination

(e) Fairness through affirmative action

(f) Reverse discrimination

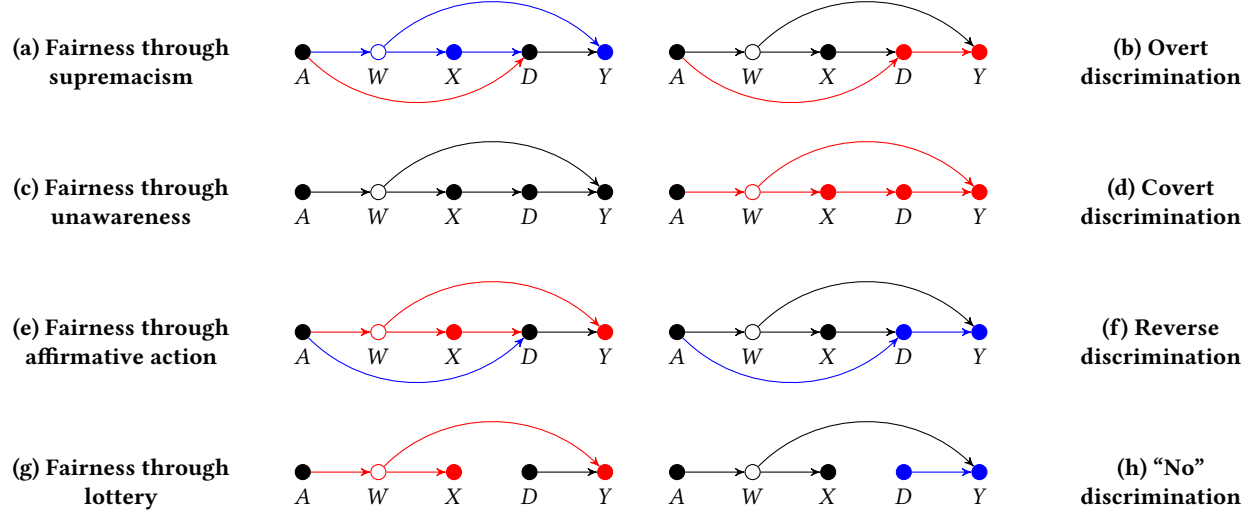**(g) Fairness through lottery**

(h) "No" discrimination

Figure 2: Modeling bias, fairness, and discrimination using causal Bayesian networks (CBNs)

To correct for covert discrimination, one might choose **fairness through affirmative action** (Figure 2e). In this form of fairness, the protected attribute A is used in the decision D ($A \rightarrow D$) to correct for bias from the data X. The intervention by the protected attribute A is colored blue, which offsets the bias from the data X that has been colored red. As a result, the final decision D is colored black to indicate that its bias has been corrected. Interestingly, the outcome Y is still colored red as the protected attribute A does not have a direct effect on the outcome Y. The protected attribute A instead has an indirect effect on the outcome Y through the decision D, but this tends to be a weaker effect. This intervention does not directly fix the underlying biases in the mediator W, which has a strong direct effect on the outcome Y ($W \rightarrow Y$). We would say that the protected attribute A has a negative indirect effect on the decision D ($A \rightarrow W \rightarrow X \rightarrow D$) and a positive direct effect on the decision D ($A \rightarrow D$), thus canceling out the biases and making the decision fair.

But if we disagree with the belief that the data is fundamentally biased (internal attribution), then fairness through affirmative action could be perceived as a form of **reverse discrimination** (Figure 2f). The intervention of the protected attribute A on the decision D ($A \rightarrow D$) could lead to individuals receiving a more positive decision D but a more negative outcome Y. For example,

a machine learning model may predict that an individual from a historically disadvantaged group may end up with a negative outcome Y. However, this would lead to a greater disparity in decisions across protected groups. So the model may instead use knowledge of that individual's membership in a protected group A to render a more positive decision D. But this positive decision D could lead to a worse outcome Y for the individual than if that individual had just received a negative decision D in the first place. An example might be approving a minority borrower for a loan even if that borrower is not creditworthy, which could result in a loan default. We would say that the protected attribute A has a positive direct effect on the decision D ($A \rightarrow D$), leading to harmful discrimination.

We now move on to the more extreme forms of fairness. Reverse discrimination could lead to **fairness through supremacism** (Figure 2a). An advantaged group may use this approach to actively suppress a disadvantaged group (e.g. white supremacism). This could happen if an advantaged group believes that a disadvantaged group has been unfairly receiving preferential treatment. For example, if a disadvantaged group has benefited from reverse discrimination in the past, then the present data X may still reflect that positive bias and is colored blue. The protected attribute A is then used to correct for that positive bias in the decision D ($A \rightarrow D$) and is colored red. This might explain what happened when hiring

managers racially discriminated against job applicants with Black sounding names compared to those with White sounding names even through the resumes were the same, as shown by Bertrand and Mullainathan[5]. We would say that the protected attribute A has a positive indirect effect on the decision D ($A \rightarrow W \rightarrow X \rightarrow D$) and a negative direct effect on the decision D ($A \rightarrow D$), thus canceling out the biases and making the decision fair.

Another extreme form of fairness is **fairness through lottery** (Figure 2g). If we believe that the data is fundamentally biased, an alternative approach to removing bias in the decision D is to remove the edge from data X to decision D. The advantage of this approach to fairness is that it provides stronger guarantees of equalizing decision rates across groups. This is particularly effective for dealing with intersectionality issues, ensuring that every potential intersection of a protected class (e.g. race, gender, and religion) receives the same decision rates. Individuals cannot change the decision by altering their own data X. The decision D could be assigned randomly or there could be a fixed rule where everyone receives a positive decision D or a negative decision D. We would say that the protected attribute A has no direct effect on the decision D and its indirect effects have been canceled out, making the decision fair.

But fairness through lottery could be interpreted as **"no" discrimination** (Figure 2h). Since data X has no effect on the decision D, then the machine learning model can literally not discriminate any differences between individuals. In doing so, it cannot cause illegal discrimination either. But this could still be interpreted as a decision that is positively biased in favor of a disadvantaged group. We would say that the protected attribute A has a positive indirect effect on the decision D ($A \rightarrow W \rightarrow X \nrightarrow D$), leading to harmful discrimination.

Note that this is not an exhaustive list of all forms of fairness and discrimination. For example, we could have a form of covert discrimination where the edges of Figure 2d are all colored in blue instead of red.

5.1 Mapping fairness to politics

Table 2 shows how each form of fairness can be mapped to different positions on the political spectrum. Each position is associated with its own fairness criterion and fairness properties.

Table 2: Mapping fairness type to the political spectrum

Political worldview	Far-left	Left-wing	Right-wing	Far-right
Type	Lottery	Affirmative action	Unaware-ness	Suprem-acism
Fairness criterion	$D \perp\!\!\!\perp A, W$	$D \perp\!\!\!\perp A$	$D \perp\!\!\!\perp A\|W$	$D \neg\!\!\perp\!\!\!\perp A$
Individual fairness	Yes	No	Yes	No
Group fairness	Yes	Yes	No	No
Pearl's law			First	Second	

Right-wing notions of fairness rely on fairness through unawareness, which is associated with the criterion $D \perp\!\!\!\perp A\|W$. That is, the decision D is conditionally independent of the protected attribute A given some mediator W. For example, a college admissions decision is independent of race given an applicant's merit. This leads to individual fairness, where similar individuals are treated similarly regardless of protected group membership[6, 8, 28, 47]. This form of fairness relies on Pearl's second fundamental law of causal inference: conditional independence.

Left-wing notions of fairness rely on affirmative action, which is associated with the criterion $D \perp\!\!\!\perp A$. That is, the decision D is independent of the protected attribute A, which is achieved by using the protected attribute A to correct for biases in the data X that is used in the decision D. We imagine what the applicant's decision would have been had they been a member of a counterfactual class. We then perform an intervention to obtain the same decision. For example, a college admissions decision would affirmatively use race to correct for deficiencies in the applicant's merit due to systemic racism. This leads to group fairness, where different groups receive more similar aggregate decision rates[6, 8, 28, 47]. This form of fairness relies on Pearl's first fundamental law of causal inference: counterfactuals and interventions.

We now examine the two extreme forms of fairness. Far-right notions of fairness rely on fairness through supremacism, which is associated with the criterion $D \neg\!\!\perp\!\!\!\perp A$. In this case, individual and group notions of fairness are both violated. Far-left notions of fairness rely on fairness through lottery, which is associated with the criterion $D \perp\!\!\!\perp A, W$. In this case, individual and group notions of fairness are both satisfied.

5.2 Mapping discrimination to antidiscrimination law

Table 3 shows how each form of discrimination can be mapped to antidiscrimination law. There are two key legal concepts in antidiscrimination law. In the US, these are called disparate treatment and disparate impact. In the EU, these are called direct discrimination and indirect discrimination. Disparate treatment or direct discrimination occurs when an individual or group is treated differently based on a protected characteristic like race or gender. Disparate impact or indirect discrimination occurs when a seemingly neutral policy has a disproportionately adverse impact on a protected group.

Table 3: Mapping discrimination type to antidiscrimination law

US Type	Disparate treatment	Disparate impact
EU Type	**Direct discrimination**	**Indirect discrimination**
Overt	X	
Covert		X
Reverse	X	
"No"		X

When a protected attribute A has a direct effect on a decision D ($A \rightarrow D$), we might consider this as a form of disparate treatment or direct discrimination. This could map to overt discrimination (Figure 2b) and reverse discrimination (Figure 2f). To avoid disparate treatment or direct discrimination, the protected attribute A should have no direct effect on the decision D.

Related to disparate treatment is the concept of redlining, where a variable that is used in a decision D acts as a proxy for a protected attribute A. For example, lenders have historically used zip code as a proxy for race as a basis for denying loans. We avoid redlining problems in this paper by only using data X that is a direct measure of the mediator W. But a deeper discussion for how to prevent redlining using CBNs is discussed by Lam[25]. He shows that certain types of alternative data can cause redlining because the protected attribute A may act as a hidden confounder, leading to that data having a spurious effect on a decision D.

When a protected attribute A has an indirect effect on a decision D, we might consider this as a form of disparate impact or indirect discrimination. This could map to covert discrimination (Figure 2d) and "no" discrimination (Figure 2h). Even though the protected attribute A has no direct effect on a decision D (making the policy neutral), the inclusion of biased data or exclusion of unbiased data could cause the protected attribute A to have an indirect effect on the decision D that adversely affects a protected group.

For disparate impact in the US, there is a three-step test[32]. The first step is to establish that a neutral policy has an adverse disparate impact on a protected group. This would involve testing whether there is a disparity in decisions D across the protected class A. The second step is to establish the business necessity of that policy. This would involve arguing that the data X being used has some causal relationship to the mediator W. The final step is to identify a less discriminatory alternative model, which would involve performing a search for a counterfactual model.

CBNs provide an analytical framework to reason about whether a model is causing illegal discrimination. By being able to explicitly model the different forms of discrimination, CBNs provide a common language that businesses and regulators can use to ensure that machine learning models are compliant with laws and regulations.

5.3 Normative effects of fairness and discrimination

CBNs are a very effective way to model the normative effects of fairness and discrimination. They allow us to encode our beliefs about how we might believe that AI systems ought to behave.

Table 4 lists the normative effects of each form of fairness as a function of how the protected attribute $A = a$ affects a decision $D = d$. The total effect is the sum of the direct and indirect effects: $TE_a(d) = DE_a(d) + IE_a(d)$. By definition, the total effect must be equal to zero for it to be considered fair: $TE_a(d) = 0$.

Table 5 lists the normative effects of each form of discrimination. By definition, the total effect must not be equal to zero for it to be considered discrimination: $TE_a(d) \neq 0$.

Notice that each form of fairness is associated with a different form of discrimination, leading to four fairness and discrimination pairs. That is, what might be considered fair for one group of people may be considered biased and discriminatory for another group of people. Fairness through supremacism (Figure 2a) and overt discrimination (Figure 2b) are both associated with the protected attribute A having a negative direct effect on the decision D ($A \rightarrow D$). Fairness through unawareness (Figure 2c) and covert discrimination (Figure 2d) are both associated with the protected attribute A having no direct effect on the decision D. Fairness through affirmative action (Figure 2e) and reverse discrimination (Figure 2f) are both associated with the protected attribute A having a positive direct effect on the decision D ($A \rightarrow D$). Finally, fairness through lottery (Figure 2g) and "no" discrimination (Figure 2h) are both associated with the protected attribute A having no direct effect on the decision D and a positive indirect effect on the decision D ($A \rightarrow W \rightarrow X \nrightarrow D$).

5.4 The limitations of causal Bayesian networks

CBNs have a key limitation: they can only be represented using directed acyclic graphs (DAGs). Recall from the block diagram in Figure 1a that outcome Y could be recorded as future data X. We can modify the CBN from Figure 1b to model this feedback loop by adding a edge from outcome Y to the mediator W as shown in Figure 3 (this edge has been colored red), which is then recorded as data X. Even though this violates the DAG constraint of CBNs, we still need to model how different fairness policies may perpetuate disparities across protected groups over time. In the next section, we will show how to overcome this limitation using causal loop diagrams.

Table 4: Normative effects of fairness

Effect	Lottery	Affirmative action	Unawareness	Supremacism
$DE_a(d)$	0	+	0	−
$IE_a(d)$	− +	−	0	+
$TE_a(d)$	0	0	0	0

Table 5: Normative effects of discrimination

Effect	"No"	Reverse	Covert	Overt
$DE_a(d)$	0	+	0	−
$IE_a(d)$	+	0	−	0
$TE_a(d)$	+	+	−	−
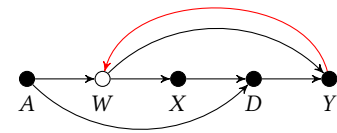
Protected   Mediator   Data   Decision   Outcome
Attribute

Figure 3: Adding a feedback loop to the causal Bayesian network

6 MODELING FAIRNESS AS A SYSTEM DYNAMICS PROBLEM

We can transform the block diagram from Figure 1a and the CBN from Figure 1b into a causal loop diagram (CLD) as shown in Figure 1c. CLDs were pioneered by Forrester [12] and are similar to CBNs in that they allow us to symbolically encode knowledge and assumptions about cause and effect relationships between variables[44]. But unlike CBNs, CLDs allow us to model feedback loops. This is useful for modeling why we might believe that disparities continue to exist between protected groups.

The CLD in Figure 1c is a simplified version of the full CLD that we will present later in this section. This shows how the same variables in the block diagram and the CBN can be mapped onto a CLD. In the center of the CLD is a node for decision D, which like before contains our machine learning model. This CLD is two sided, with the left side representing group A and the right side representing group B. Data from group A X_A and group B X_B is fed into the ML model, which renders a decision D ($X_A \rightarrow D \leftarrow X_B$). Note that there is a time delay (denoted by ‖) between decision D and outcomes Y_A and Y_B. The outcomes Y_A and Y_B are recorded as future data X_A and X_B ($Y_A \rightarrow X_A$ and $Y_B \rightarrow X_B$), which in turn drives a future decision D ($X_A \rightarrow D \leftarrow X_B$). The outcomes Y_A and Y_B also influence whether the protected attribute A_A or A_B might be used to influence a decision D. For example, if group B has historically received worse outcomes that group A, then the group might demand the protected attribute A_B to be affirmatively used to modify decision D (fairness through affirmative action). However, the use of the protected attribute A_A or A_B could also impact outcomes Y_A or Y_B.

Within a CLD, every path is labeled as having a positive (+) or negative (−) direction. If the path is positive, then changes in one variable leads to changes in another variable in the same direction. For example, an increase in one variable would lead to an increase in the other variable. Alternatively, a decrease in one variable would lead to a decrease in the other variable. If the path is negative, then changes in one variable leads to changes in another variable in the opposite direction. For example, an increase in one variable would lead to a decrease in the other variable. Alternatively, a decrease in one variable would lead to an increase in the other variable.

CLDs are represented using a combination of reinforcing loops or balancing loops. Reinforcing loops compound changes over time, creating either virtuous cycles that cause growth or vicious cycles that cause decay. Balancing loops counteract changes over time, keeping a system within a fixed state that prevents growth or decay. Reinforcing loops have an even number of negative paths, while balancing loops have an odd number of negative paths.

We can combine reinforcing and balancing loops to create system archetypes. These represent common recurring patterns that explain counterintuitive behavior in complex systems. This concept was first coined by Senge[41] with additional research done by Meadows[27] and Sterman[44]. Three system archetypes seem particularly helpful for understanding the fairness problem, which we have adapted from Kim[22].

The first archetype that we will examine is the **"success to the successful"** archetype, which is shown in Figure 4. This archetype consists of two reinforcing loops, R1 and R2. Let's say that we

have two equally capable groups that are competing for the same resources: groups A and B. If group A was initially given more resources, then that group will become more successful. That success provides justification to allocate even more resources to group A. This is shown by reinforcing loop R1. However, that could mean fewer resources are allocated to group B, which causes that group to become less successful. This results in fewer resources being allocated to group B. This is shown by reinforcing loop R2. Over the long run, R1 could create a virtuous cycle of growth for group A, while R2 could create a vicious cycle of decay for group B.
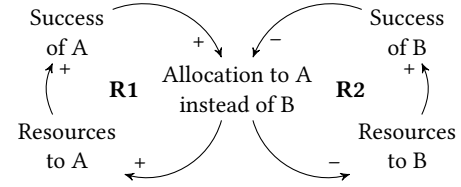

Figure 4: Success to the successful archetype

The next archetype that we will examine is the **"limits to success"** archetype, which is shown in Figure 5. This archetype consists of a reinforcing loop R1 and a balancing loop B1. Let's say that a system's performance continues to increase as a result of past efforts. This is shown by the reinforcing loop R1. But as performance increases, it begins to plateau as it reaches some limiting action. The limiting action is governed by some constraint that prevents further performance improvement. This is shown by the balancing loop B1. Even as efforts increase, the limiting action acts as a ceiling on performance.
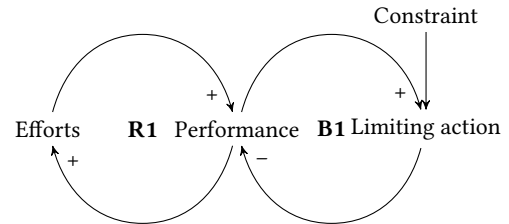

Figure 5: Limits to success archetype

The final system archetype that we will examine is the **"shifting the burden / addiction"** archetype, which is shown in Figure 6. Let's say that there is a problem symptom and two choices for a solution: an internal solution and an external intervention. The internal solution takes a long time to correct for the problem symptom (the double hash marks ‖ between the internal solution and the problem symptom represent a time delay), meaning that the problem symptom would continue to persist in the short term. On the other hand, the external intervention might alleviate the problem symptom immediately. But an unintended side effect of this solution is that it could create a dependence on the external intervention, which diverts attention away from an internal solution. This means that the problem symptom would actually persist over the long term, offsetting the benefit of the external intervention. This is called

the shifting the burden archetype because the burden of solving the problem symptom is shifted from one party to another. It is also called an addiction archetype because people who rely on an external intervention may become addicted to that solution over the long term.
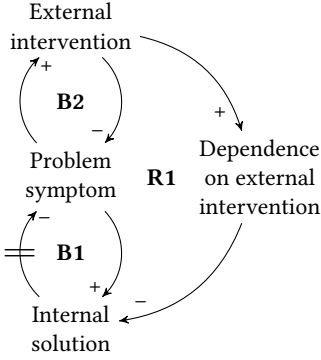

External
intervention
+
B2
−
Problem   **R1**   Dependence
symptom   on external
+   intervention
−
B1
+
Internal   −
solution

Figure 6: Shifting the burden / addiction archetype

In Figure 7, we combine these three archetypes into a single CLD to model the algorithmic fairness problem. This diagram is an expanded version of the CLD in Figure 1c. We begin with the "success to the successful" archetype. A machine learning model is embedded inside decision D, which makes allocation decisions based on decision X and has a tendency to favor one group over another. Those decisions lead to resources being assigned to each group, which eventually determines that group's success and is measured as an outcome Y. Past success Y becomes future data X, which is then fed back into the machine learning model to drive a future allocation decision D. This is represented using reinforcing loops R1 and R2. Since the allocation decisions D are initially only driven by data X, success to the successful is equivalent to fairness through unawareness (Figure 2c).

We next add in the "limits to success" archetype. As resources to group A or group B increases, it may hit up against a limiting action that is governed by come constraint. This then reduces the success of that group as measured by outcome Y. As the group's success becomes limited, this is reflected in the data X for that group. This leads to a reduction in the allocation decision D of resources to the previously successful group. This is represented by balancing loops B1 and B2. Whereas success to the successful causes a previously successful group to become more successful, limits to success may give a less successful group a chance to catch up.

Finally, we add in the "shifting the burden / addiction" archetype. If one group is allocated less resources than another group, then it could breed resentment by the disadvantaged group. That group could then demand that the protected attribute A be used to externally intervene on the allocation decision D. This external intervention is equivalent to fairness through affirmative action (Figure 2e). While it might reduce allocation disparities in the short term, it may actually create a dependence on the external intervention over the long term. This is because if allocation decisions D are not based purely on data X, which is a measure of mediator W (e.g. merit, qualifications), then the assignment of resources to the

disadvantaged group may not lead to their success Y. For example, approving a loan to a minority applicant even though the applicant is not creditworthy could lead to a worse outcome (default) than if the loan was denied in the first place. That decreased success in outcome Y for the less successful group could trigger a demand for an even stronger external intervention A on the decision D, reinforcing this problem. In the meanwhile, attention may be taken away from the internal solution of optimizing for the mediator W and the data X so that the disadvantaged group would perform better under fairness through unawareness. This may especially be the case if the disadvantaged group believes that the mediator W is biased and therefore does not focus on optimizing that metric.

6.1 Positive effects of fairness and discrimination

Whereas CBNs are very effective at modeling the normative effects of fairness and discrimination, CLDs are very effective at modeling the positive effects. Recall from Hume's law that positive effects model what is, while normative effects model what ought to be (i.e. the is–ought problem).

Table 6 lists the positive effects of the different forms of fairness and discrimination. As mentioned before, each form of fairness is associated with a form of discrimination, leading to the same direct effect $DE_a(d)$. Whereas normative effects differ based on our prior beliefs about where bias exists in a model, positive effects do not take into account our prior beliefs of bias. However our analysis of positive effects depends on whether we analyze disparities across groups from a short-term or a long-term perspective. Notice that for fairness through supremacism or overt discrimination, since a disadvantaged group is being overtly disadvantaged, there is a doubly negative total effect ($TE_a(d) \ll 0$).

Table 6: Positive effects of fairness and discrimination

Fairness	Lottery	Affirmative action	Unaware-ness	Suprema-cism
Discrim-ination	"No"	Reverse	Covert	Overt
Short-term				
$DE_a(d)$	0	+	0	−
$IE_a(d)$	0	−	−	−
$TE_a(d)$	0	0	−	− −
Long-term				
$DE_a(d)$	0	+	0	−
$IE_a(d)$	0	− −	− +	−
$TE_a(d)$	0	−	0	− −

During the short-term, we only examine the "success to the successful" archetype as shown in Figure 7 through feedback loops R1 and R2. A policy of fairness through unawareness would lead to a disadvantaged group receiving worse decisions than an advantaged group ($TE_a(d) < 0$). This is because fairness through unawareness would replicate and reinforce the past success of the advantaged group over the disadvantaged group ($IE_a(d) < 0$). But
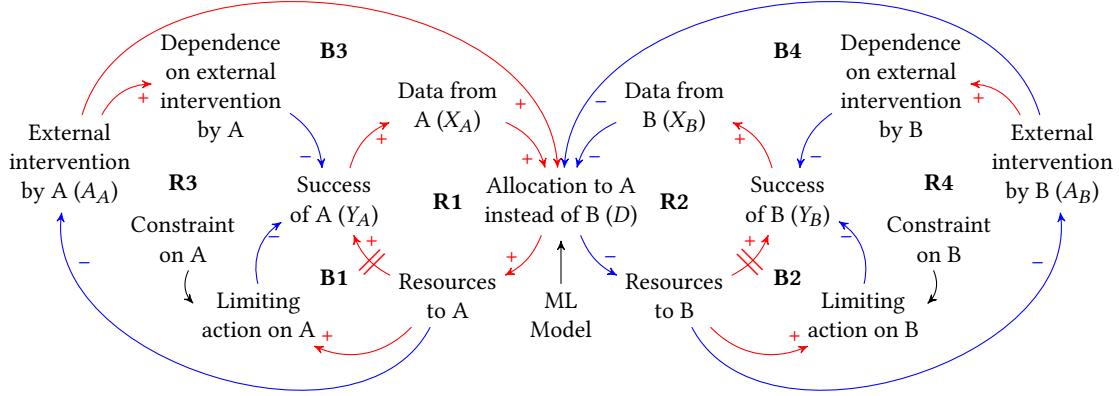
Figure 7: Causal loop diagram representation of algorithmic fairness

an affirmative action policy could help to equalize the decision rates across both groups. Therefore, the total effect from fairness through unawareness is negative ($TE_a(d) < 0$) and the total effect from affirmative action is zero ($TE_a(d) = 0$).

During the long-term, we add in the effects of the "limits to success" archetype as shown in Figure 7 through feedback loops B1 and B2. This archetype puts a constraint on how successful the advantaged group can become, reducing the disparities between the advantaged group and the disadvantaged group ($IE_a(d) > 0$). We also add in the effects of the "shifting the burden / addiction" archetype as shown in Figure 7 through feedback loops R3, B3, R4, and B4. This archetype can create a dependency by the disadvantaged group on the external intervention and harm their long-term success ($IE_a(d) < 0$). Therefore, the total effect from fairness through unawareness is zero ($TE_a(d) = 0$) and the total effect from affirmative action is negative ($TE_a(d) < 0$). That is, the total effects between the short-term and long-term versions of these fairness policies may flip direction.

6.2 Case study: Racial disparities in the United States

We can use the causal graphs from Figure 2 to demonstrate how to model the complex political debate over racial disparities in the US. According to the US Census Bureau, median household income in 2021 was $100,573 for Asians, $74,932 for Whites, $57,671 for Hispanics, and $48,297 for Blacks[45]. Meanwhile, median household wealth in 2021 was $320,900 for Asians, $250,400 for Whites, $48,700 for Hispanics, and $27,100 for Blacks[23]. We would like to understand why Whites and Asians have higher median household income and wealth than Blacks and Hispanics.

When it comes to understanding the root causes of racial disparities, we can take into account some of the original sources of bias in the data. We know that the US has historically had unjust policies that overtly discriminated against non-Whites. These included overtly racist policies like slavery, Jim Crow segregation, and the

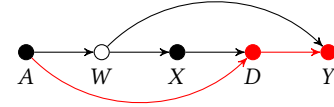Chinese Exclusion Act. For convenience, we replicate the CBN for overt discrimination in Figure 8.

Figure 8: Overt discrimination

Overt discrimination causes bias in both the decision D and the outcome Y. The outcome Y becomes future data X due to a feedback loop. Figure 9 shows what this feedback loop looks like within a CBN. We temporarily violate the directed acyclic graph constraint of CBNs and create a path $Y \rightarrow W$, causing the mediator W to become biased.
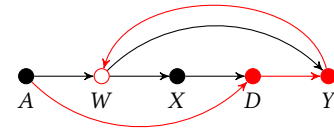

Figure 9: Feedback from overt discrimination

The bias in mediator W gets replicated in the data X, which then causes bias in future decisions D and outcomes Y. Even though the US passed the Civil Rights Act of 1964 banning overt discrimination, this didn't reverse the initial bias caused by historically unjust policies against non-Whites. Even a racially neutral policy could still form a basis for covert discrimination as shown in Figure 10.

This creates a dilemma for policymakers: Should the US institute fairness through unawareness, even if such a policy could replicate past racial disparities (i.e. the "success to the successful" archetype
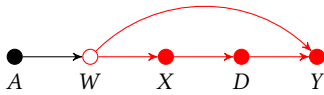
Figure 10: Covert discrimination

as shown in Figure 4)? Or should the US institute affirmative action to try to correct for past injustices (i.e. the "shifting the burden" archetype as shown in Figure 6)?

Due to the "success to the successful" archetype, those racial groups that have traditionally been left behind may not be able to compete against other racial groups that are presently more successful. For example, Whites and Asians have access to better funded public schools, safer neighborhoods, higher quality healthcare, and better social networks for job opportunities. These advantages could result in Whites and Asians being able to outcompete Blacks and Hispanics for the same resources under fairness through unawareness, since there will be a disparity in their merit and qualifications.

So let's say that US policymakers (particularly Democrats) succeed in instituting affirmative action as shown in Figure 11. While affirmative action might help correct for biases in the decision D, notice in Figure 11 that it does not necessarily correct for biases in the outcome Y. The protected attribute A does not have a direct effect on the outcome Y. It can only directly affect the decision D. Correcting for biases in the decision D cannot fully offset biases in the mediator W, which means that the outcome Y remains biased.
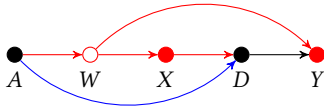

Figure 11: Fairness through affirmative action

The path $D \rightarrow Y$ only has a weak effect on the outcome Y. On the other hand, the path $W \rightarrow Y$ has a strong effect on the outcome Y. This is shown in Figure 12.
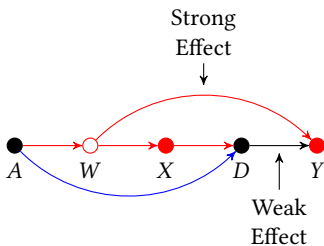

Figure 12: Strong and weak effects

For example, an affirmative action policy in college admissions might help more Blacks and Hispanics to be accepted ($A \rightarrow D$). But because they were not accepted purely based on their merit (mediator W), they might be less likely to graduate (outcome Y). This is because they might be put into a class where their White and

Asian counterparts would consistently outperform them. Another example could be an affirmative action policy for credit decisions. This might help more Blacks and Hispanics to receive more loans ($A \rightarrow D$). But since they were not accepted purely based on their creditworthiness (mediator W), they would be more likely to default (outcome Y). This is because affirmative action does not directly improve one's merit or creditworthiness. Even worse, affirmative action could have the opposite effect on the mediator W.

Due to the "shifting the burden / addiction" archetype, an affirmative action policy may reinforce a dependency on this external intervention and take away attention from an internal solution. Figure 13 shows that the internal solution would be to focus on reducing disparities in the mediator W. But instead, affirmative action shifts attention to reducing disparities in the decision D.
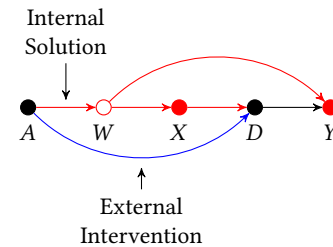

Figure 13: Internal solution and external intervention

This creates a major problem. Affirmative action may create a false illusion that you can have a system of double standards and not pay the price. While it could enable less qualified Blacks and Hispanics to get access to the some of the same resources that would have went to more qualified Whites and Asians, it doesn't necessarily translate into better overall outcomes. In college admissions for example, one could argue that a Black or Hispanic college student struggling at an elite university might have been better off thriving at a flagship state university instead. In employment, a Black or Hispanic employee might have been better off being placed in a team where they were at the same level as their White and Asian peers versus being placed in a higher performing team where they ended up dragging the rest of the team down. This is the main argument that has been advanced by Sowell, Mac Donald, and other conservative scholars[10, 42, 43]. This is called mismatch theory by Sander [39].

Placing Blacks and Hispanics into higher level positions than is warranted by their merits and qualifications could also exact a harmful psychological toll. This could reinforce their views that they are victims of an unjust system[18], leading to a sense of learned helplessness[40] and a loss of agency. This could result in a reduction of human capital and productivity, which is the primary determinant of income and wealth in the primarily capitalist economic system of the US[4]. Rather than helping Blacks and Hispanics become more competitive against Whites and Asians, affirmative action may have instead made them less competitive.

Asians in the US have not historically benefited from affirmative action policies. Asians have instead relied on fairness through unawareness, as shown in Figure 14. In some cases like college

admissions, they may even have to outperform their White counterparts to receive the same admission decisions[13]. Asians as a group have tended to focus their efforts on the internal solution of competing based on their merits and qualifications. By working against Asians, affirmative action programs may have actually caused Asians to work even harder than other racial groups. By focusing on developing their own human capital instead of demanding external government interventions, not only have Asians closed the gap in income and wealth with Whites over the long-term, they have actually surpassed them. Over time, this has also reduced perceived biases in the data between Asians and Whites.
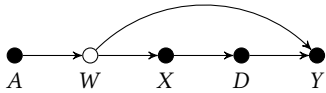

Figure 14: Fairness through unawareness

Due to the "limits to success" archetype, human capital represents a constraint on the level of success for any racial group. By focusing more on education and pursuing higher income careers (e.g. STEM) than other racial groups, Asians have been able to catch up and surpass Whites in terms of income and wealth. This may reflect a selection bias in the types of Asians who emigrate to the US, who oftentimes come to pursue better educational and career opportunities.

This is the reason why some US policymakers (particularly Republicans) have been arguing in favor of fairness through unawareness. While acknowledging that historically unjust policies like slavery and Jim Crow have created the initial disparities between Whites and non-Whites, they also recognize that external government interventions like affirmative action may cause more harm than good. They would argue that it is not an accident that the racial group that has benefited the most from affirmative action (i.e. Blacks) ultimately received the worst outcomes. The focus by the Black community on how past injustices perpetuate racial disparities may create a self-fulfilling prophecy. Meanwhile, the racial group that has benefited the least from affirmative action (i.e. Asians) ultimately received the best outcomes. The focus by the Asian community on developing their own human capital through education and hard work (i.e. the model minority) has allowed that community to close their initial disparities with Whites.

On the other hand, some Democratic policymakers have argued for a doubling down on affirmative action and DEI programs. This was especially the case during the past four years under the Biden administration. To address the fact that affirmative action only has a weak effect on outcomes, their goal was to deploy affirmative action and DEI programs across the entire US economy. This would have created a racially segregated economic system that provides special treatment for certain minority groups. By instituting a system of double standards and quotas across the US economy, their hope was to equalize decisions and outcomes across the racial groups. Doing so would have required a radical restructuring of the US economy, from one that primarily allocates resources based on merit to one that primarily allocates resources based on equity. It would have reduced the size of the overall economic pie as opportunities are distributed not primarily based on qualifications but instead to reach diversity quotas. As we saw in the recent election, there was not enough popular support in the US for such a radical transformation of the US economy.

That being said, this doesn't mean that affirmative action would not have any value. If it can be used to identify and correct for false positives, then affirmative action might be a helpful way to marginally reduce disparities across groups. The key here is to use the protected attribute A to only correct for biases in the data X that may overpredict the probability of a bad outcome Y, as opposed to using it to directly achieve some type of diversity quota. We would need to ensure that an affirmative action policy that reduces disparities in the decision D does not result in an increase in the disparities in the outcome Y.

Affirmative action is a low-leverage solution that at best can only modestly decrease racial disparities. It is not a fundamental, high-leverage solution for reducing racial disparities over the long-term. As long as there are racial disparities in the mediator W, then there will be racial disparities in the outcome Y. This would mean that racial disparities in household income and wealth in the US would continue to persist. The focus should instead be on closing disparities in the mediator W. For example, Democrats could call for increased government funding of education, training, and health programs to help minorities improve their qualifications so that they can be more productive. On the other hand, Republicans would continue their calls for the banning of racial preferences in high-stakes decisions to keep all racial groups accountable to the same high standards.

These models provide a basis for acknowledging that while historical policies like slavery and Jim Crow generated the initial racial disparities between Whites and Blacks, contemporary policies like affirmative action and DEI may have reinforced those disparities. Affirmative action should at most be used as a temporary solution for marginally closing the racial equity gap. If it is used, then there should be a long-term plan to eventually switch back to fairness through unawareness. If an addictive dependency on affirmative action develops that diverts attention away from internal solutions for closing the merit and qualifications gap, then the affirmative action policy should be ended. Systems thinking shows us that AI researchers should not have completely dismissed fairness through unawareness as a fundamental solution for closing the racial equity gap.

7 MODELING FAIRNESS AS A CAUSAL HIERARCHY

Systems thinking allows us to combine three previously independent analytical techniques: machine learning, causal inference, and system dynamics. Doing so allows us to capture the emergent behavior from three different aspects of the fairness problem. We can think of CBNs as a higher level representation of supervised machine learning models, which emerges when we "think outside of the black box." We can think of CLDs as a higher level representation of CBNs, which emerges when we violate the directed acyclic graph constraint and create a feedback loop. Another way to think of this is that a CBN may act as a "wrapper" around a supervised ML model and as a "snapshot" of a CLD.

Layer 3. Counterfactuals	Layer 3. Dynamic causation
Activity: Imagining, retrospection, understanding	**Field:** System dynamics (Jay Forrester)
Question: What if I had done...? Why?	**Representation:** Causal loop diagrams
Example: Was it the aspirin that stopped my headache?	**Application:** Simulation

<div align="center">↑ Violate the DAG constraint! ↑</div>

Layer 2. Intervention	Layer 2. Static causation
Activity: Doing, intervening	**Field:** Causal inference (Judea Pearl)
Question: What if I do...? How?	**Representation:** Causal Bayesian networks
Example: If I take aspirin, will my headache be cured?	**Application:** Control

<div align="center">↑ Think outside the black box! ↑</div>

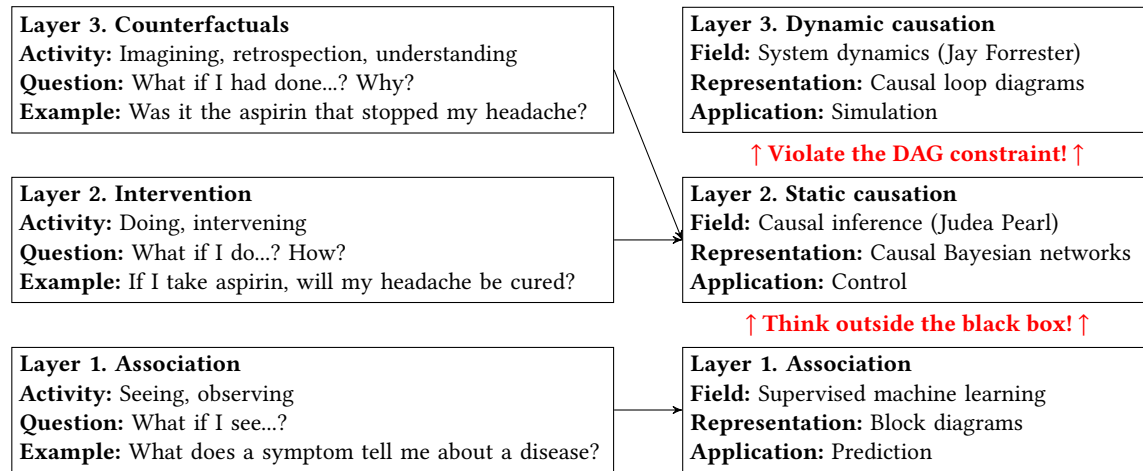| Layer 1. Association	Layer 1. Association
Activity: Seeing, observing	**Field:** Supervised machine learning
Question: What if I see...?	**Representation:** Block diagrams
Example: What does a symptom tell me about a disease?	**Application:** Prediction

Figure 15: Causal hierarchy representation of algorithmic fairness

However, this suggests the need for a new causal hierarchy to model the algorithmic fairness problem. The left side of Figure 15 shows the causal hierarchy proposed by Pearl, which is not optimized for supervised machine learning and does not incorporate system dynamics. The right side of Figure 15 is a proposed new causal hierarchy that more explicitly models the emergent behavior between supervised machine learning, causal inference, and system dynamics.

We can directly map layer 1 of Pearl's causal hierarchy (association) to layer 1 of this new causal hierarchy (association), but we explicitly define this layer to be used for supervised machine learning. We can then combine layers 2 and 3 of Pearl's causal hierarchy (intervention and counterfactuals) into layer 2 of this new causal hierarchy (acyclic causation). Finally, we add a new layer 3 to the causal hierarchy for system dynamics (cyclic causation).

Each of the layers in this new causal hierarchy have unique properties and behaviors that make them optimized for different applications. The first layer is associated with supervised machine learning and is optimized for prediction. The second layer is associated with causal inference and is optimized for control. The third layer is associated with system dynamics and is optimized for simulation. These three new causal layers make up an expanded causal hierarchy that is consistent with systems thinking principles.

8 MODELING FAIRNESS AS A SYSTEMS MAP

We can create an even higher level representation of the algorithmic fairness problem called a systems map. This is shown in Figure 16. On the left side of the systems map, we use a connection circle to show the different fields in the social sciences that are related to fairness. These ideas were described in greater detail in Section 3. On the right side of the systems map, we use a causal hierarchy to show the different technical aspects of the fairness problem. These ideas were described in greater detail in Section 7.

As mentioned before, algorithmic fairness is a sociotechnical problem. We need to find a way to bridge the social and technical aspects of the fairness problem. In each of the social science fields, there exists a dichotomy that represents the two sides of the

fairness problem. This allows us to understand the fundamentally ideological and political debate over fairness, which centers around how we attribute the causes for disparities across protected groups. The systems map shows a causal bridge that links this dualism in the social sciences to Pearl's two fundamental laws of causal inference. This systems map provides a very compact and high level representation for how the different fields of fairness may relate to each other.

9 CONCLUSION

Systems thinking provides us with a novel way to model the algorithmic fairness problem by "thinking outside the black box." While a large amount of AI research has been focused on the hidden layers that are "inside" a black box model, this paper has focused on the causal layers that are "outside" a black box model. These outer layers are where we can encode our knowledge and assumptions about where we might believe that bias occurs in the data generating process. We showed that the first outer layer can be represented using causal Bayesian networks and the second outer layer can be represented using causal loop diagrams. This gives us an elegant way of modeling the emergent behavior between machine learning, causal inference, and system dynamics.

This approach allow us to develop a more unified and holistic view of the fairness and machine learning problem, providing essential connections between computer science, politics, and the law. This is the key to building AI systems that are aligned with society's shared democratic values and the policymakers' political agendas. It will enable businesses and regulators to communicate with each other using a common causal language, one that can be used by both data scientists and lawyers. These models may help to identify mathematical flaws in the existing laws and regulations. They could help businesses to comply with antidiscrimination laws and regulations, reducing business uncertainty and expensive litigation while providing a much needed layer of transparency and trust for consumers of AI products and services.

Systems thinking supplies the missing mathematical models that are needed to build responsible AI systems. To solve fairness, we
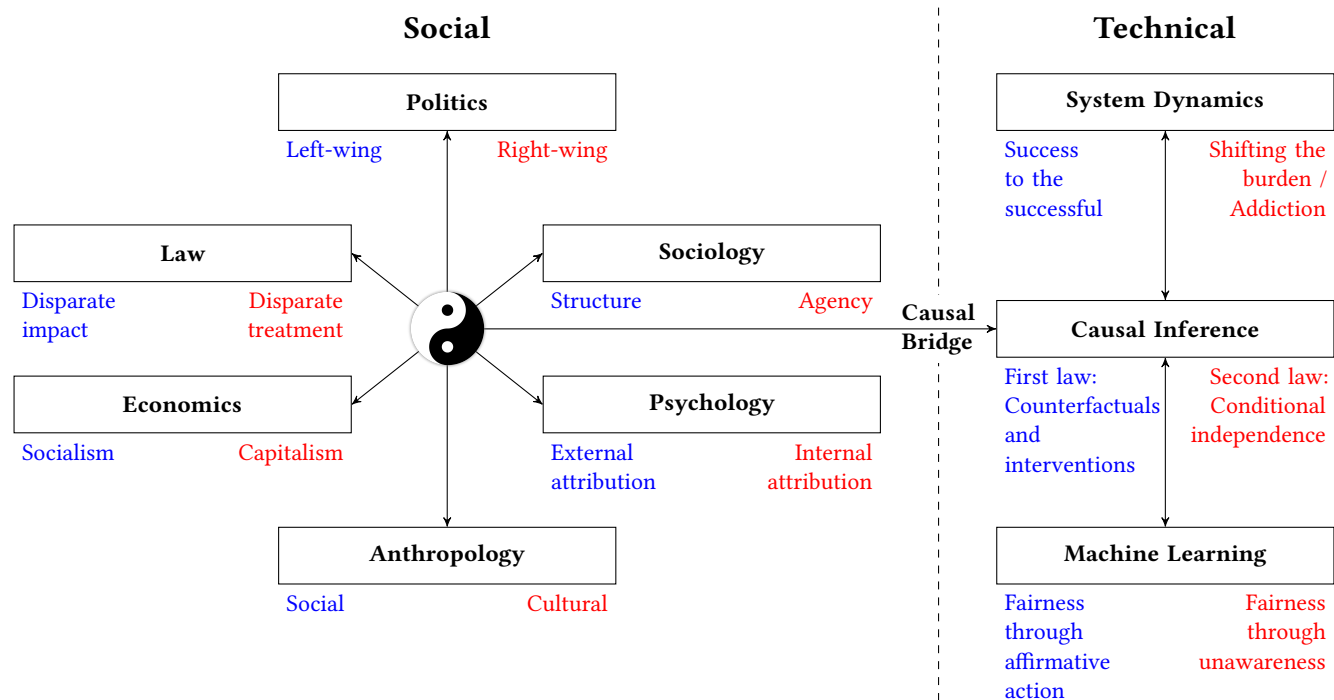
Figure 16: Systems map representation of algorithmic fairness

need to bring together both the social and technical aspects of what is ultimately a complex systems problem. On the social side, it requires modeling the diversity of perspectives that come from both sides of the political spectrum. On the technical side, it requires unifying analytical techniques that were previously thought of as being independent. By unifying all of these concepts together, we hope that this enables the development of AI policies and systems that could lead to pragmatic, real-world solutions for reducing disparities across protected groups in society.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. Pro Publica Inc. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Chris Barker. 2005. *Cultural Studies: Theory and Practice*. Sage, London, UK.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MA.

[4] Gary S. Becker. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. The University of Chicago Press, London, UK.

[5] Marianne Bertrand and Sendhil Mullainathan. 2003. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*. Working Paper 9873. National Bureau of Economic Research. https://doi.org/10.3386/w9873

[6] Reuben Binns. 2019. On the Apparent Conflict Between Individual and Group Fairness. arXiv:1912.06883 [cs.LG] https://arxiv.org/abs/1912.06883

[7] Paul Carus. 1910. Person and personality. *The Monist* 20 (1910), 379.

[8] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 1–21.

[9] Silvia Chiappa and William S. Isaac. 2019. *A Causal Bayesian Networks Viewpoint on Fairness*. Springer International Publishing, Cham, Switzerland, 3–20. https://doi.org/10.1007/978-3-030-16744-8_1

[10] Heather Mac Donald. 2023. *When Race Trumps Merit: How the Pursuit of Equity Sacrifices Excellence, Destroys Beauty, and Threatens Lives*. DW Books, Nashville, TN.

[11] Carol S. Dweck. 2007. *Mindset: The New Psychology of Success*. Ballantine Books, New York.

[12] Jay Forrester. 1971. Counterintuitive behavior of social systems. *Technology Review* 53, 3 (1971), 52–68.

[13] Joshua Grossman, Sabina Tomkins, Lindsay C Page, and Sharad Goel. 2023. *The Disparate Impacts of College Admissions Policies on Asian American Applicants*. Working Paper 31527. National Bureau of Economic Research. https://doi.org/10.3386/w31527

[14] Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. Lawrence Erlbaum Associates, Hillsdale, NJ.

[15] Donald Martin Jr. 2021. *Applying Community-Based System Dynamics to Combat AI Bias*. System Dynamics Society. https://www.youtube.com/watch?v=LpOZXYlQOFw

[16] Donald Martin Jr. 2023. *Using societal context knowledge to foster the responsible application of AI*. Google Research. https://research.google/blog/using-societal-context-knowledge-to-foster-the-responsible-application-of-ai/

[17] Shelly Kagan. 2012. Moral Desert. In *The Geometry of Desert*. Oxford University Press, New York, NY. https://doi.org/10.1093/acprof:oso/9780199895595.003.0001 arXiv:https://academic.oup.com/book/0/chapter/150310098/chapter-ag-pdf/44982224/book_6467_section_150310098.ag.pdf

[18] Scott Barry Kaufman. 2020. Unraveling the Mindset of Victimhood. *Scientific American* 30, 5 (2020), 14.

[19] H. H. Kelley. 1967. Attribution Theory in Social Psychology. *Nebraska Symposium on Motivation* 15 (1967), 192–238.

[20] Ibram X. Kendi. 2017. *Stamped from the Beginning: The Definitive History of Racist Ideas in America*. Bold Type Books, New York, NY.

[21] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 656–666.

[22] Daniel H. Kim. 1992. *Systems Archetypes I: Diagnosing Systemic Issues and Designing High-Leverage Interventions*. Pegasus Communications, Inc, Waltham, MA.

[23] Rakesh Kochnar and Mohamad Moslimani. 2023. *Wealth Surged in the Pandemic, but Debt Endures for Poorer Black and Hispanic Families*. Technical Report. Pew Research Center.

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Red Hook, NY. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[25] Chris Lam. 2024. Debiasing Alternative Data for Credit Underwriting Using Causal Inference. arXiv:2410.22382 [q-fin.RM] https://arxiv.org/abs/2410.22382

[26] Glenn C. Loury. 2002. *The Anatomy of Racial Inequality*. Harvard University Press, Cambridge, MA.

[27] Donella H. Meadows. 2008. *Thinking in Systems*. Chelsea Green Publishing Company, White River Junction, VT.

[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. https://doi.org/10.1145/3457607

[29] Charles W. Mills. 1997. *The Racial Contract*. Cornell University Press, Ithaca, NY.

[30] Arvind Narayanan. 2018. *21 fairness definitions and their politics*. ACM Conference on Fairness, Accountability, and Transparency. https://www.youtube.com/embed/jIXIuYdnyyk

[31] Hamed Nilforoshan, Johann Gaebler, Ravi Shroff, and Sharad Goel. 2022. Causal Conceptions of Fairness and their Consequences. arXiv:2207.05302 [cs.LG] https://arxiv.org/abs/2207.05302

[32] US Department of Justice Civil Rights Division. 2024. *Title VI Legal Manual*. US Department of Justice, Washington, DC.

[33] Judea Pearl. 2000. *Causality*. Cambridge University Press, Cambridge, UK.

[34] Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Seattle, Washington) *(UAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 411–420.

[35] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics*. Wiley, West Sussex, UK.

[36] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, NY.

[37] Drago Plecko and Elias Bareinboim. 2022. Causal Fairness Analysis. arXiv:2207.11385 [cs.AI] https://arxiv.org/abs/2207.11385

[38] Julian B Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* 80, 1 (1966), 1–28.

[39] Richard H. Sander and Stuart Taylor. 2012. *Mismatch: How Affirmative Action Hurts Students It's Intended to Help, and Why Universities Won't Admit It*. Basic Books, New York.

[40] M E P Seligman. 1972. Learned Helplessness. *Annual Review of Medicine* 23, Volume 23, 1972 (1972), 407–412. https://doi.org/10.1146/annurev.me.23.020172.002203

[41] Peter M. Senge. 1990. *The Fifth Discipline: The Art and Practice of The Learning Organization*. Doubleday Publishing Group, New York, NY.

[42] Thomas Sowell. 2004. *Affirmative Action Around the World: An Empirical Study*. Yale University Press, New Haven, CT.

[43] Thomas Sowell. 2019. *Discrimination and Disparities*. Basic Books, New York, NY.

[44] John D. Sterman. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw-Hill, Boston, MA.

[45] American Community Survey. 2021. *S0201: Selected Population Profile in the United States*. Technical Report. United States Census Bureau.

[46] Matthew Talbert. 2024. Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (Fall 2024 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University, Stanford, CA.

[47] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) *(FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[48] Bernard Weiner. 1986. *An Attributional Theory of Motivation and Emotion*. Springer, New York. NY.