RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity

David Liu*
Northeastern University
Boston, MA, USA
liu.davi@northeastern.edu

Zohair Shafi*
Northeastern University
Boston, MA, USA
shafi.z@northeastern.edu

William Fleisher
Northeastern University
Boston, MA, USA
w.fleisher@northeastern.edu

Tina Eliassi-Rad
Northeastern University
Boston, USA
tina@eliassi.org

Scott Alfeld
Amherst College
Amherst, MA, USA
salfeld@amherst.edu

ABSTRACT

We present RAWLSNET, a system for altering Bayesian Network (BN) models to satisfy the Rawlsian principle of *fair equality of opportunity* (FEO). RAWLSNET's BN models generate aspirational data distributions: data generated to reflect an ideally fair, FEO-satisfying society. FEO states that everyone with the same talent and willingness to use it should have the same chance of achieving advantageous social positions (e.g., employment), regardless of their background circumstances (e.g., socioeconomic status). Satisfying FEO requires alterations to social structures such as school assignments. Our paper describes RAWLSNET, a method which takes as input a BN representation of an FEO application and alters the BN's parameters so as to satisfy FEO when possible, and minimize deviation from FEO otherwise. We also offer guidance for applying RAWLSNET, including on recognizing proper applications of FEO. We demonstrate the use of RAWLSNET with publicly available data sets. RAWLSNET's altered BNs offer the novel capability of generating *aspirational data* for FEO-relevant tasks. Aspirational data are free from biases of real-world data, and thus are useful for recognizing and detecting sources of unfairness in machine learning algorithms besides biased data.

CCS CONCEPTS

• **Mathematics of computing** → **Bayesian networks**; • **Computing methodologies** → **Bayesian network models**; • **Applied computing** → **Sociology**.

KEYWORDS

Rawlsian fair equality of opportunity, Bayesian networks, aspirational data

*Both authors contributed equally to this research.

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462618

ACM Reference Format:
David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. 2021. RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462618

1 INTRODUCTION

Machine learning algorithms often display pernicious biases that lead to harmful and unfair outcomes for marginalized groups [4, 9, 15, 17, 38, 39]. In response to this algorithmic bias, there is a widely growing literature seeking to achieve fairness in machine learning [3, 25, 40]. As a guide to fairness, we appeal to John Rawls' theory of *justice as fairness* [42, 43]. For a society to be fair in the Rawlsian sense, it must satisfy a substantive equality of opportunity principle, which Rawls calls *fair equality of opportunity* (FEO). This principle governs fair decision-making in the context of distributing desirable social positions (such as employment) in society. Specifically, it requires that all people who have the same level of talent and ambition have the same chance of attaining advantageous social positions. We address the following question: *Given an unfair (in the Rawlsian sense) outcome and the capability to alter some (but not all) decision-making processes, how can one satisfy FEO?* To this end, we use Bayesian Networks (BNs) to model decisions that are governed by FEO — namely, those that impact the distribution of advantageous social positions. We then give a characterization of FEO in terms of conditional probabilities, which allows for the mathematical formalization of the above problem. Specifically, we present RAWLSNET. Given data, RAWLSNET offers guidance on making decisions that satisfy FEO. When satisfying FEO is impossible (e.g., due to resource constraints such as number of available jobs), RAWLSNET finds the "closest" solution to satisfying FEO. See Section 3 for details on "closest."

RAWLSNET has the following three components: (1) learn a BN; (2) determine relevance to FEO; (3) update parameters of the learned BN to satisfy FEO if possible. Otherwise, update the parameters to approximately satisfy FEO. To learn the BN, RAWLSNET accepts as input (*i*) a fully trained Bayesian Network (BN), or (*ii*) the structure of a BN and data to learn the BN's parameters (i.e., probabilities in its conditional probability tables – a.k.a. CPTs), or (*iii*) data to learn the BN's structure and its parameters. If RAWLSNET has to learn the BN structure, it asks the user to answer a series of queries:

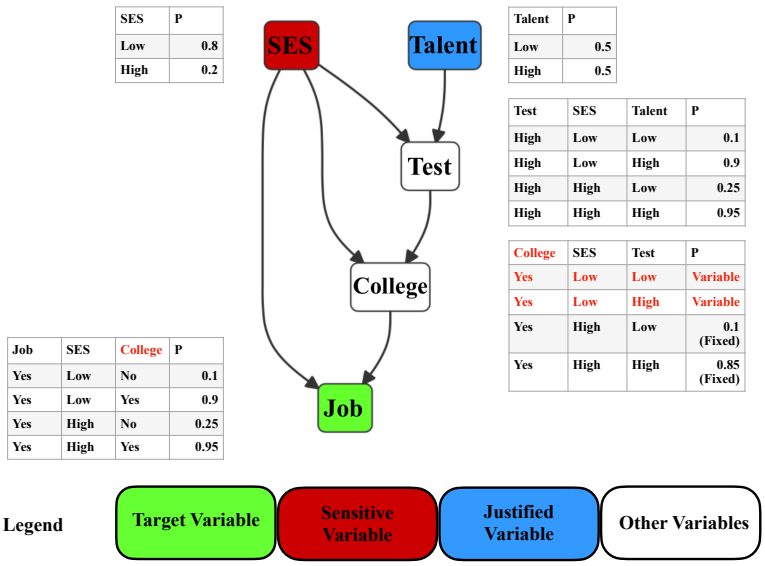


Figure 1: Running example of a Bayesian Network for college admissions. The Talent variable refers to having the innate capability to succeed at a job. The Socioeconomic Status (SES) variable represents socioeconomic status (e.g., an individual with a high SES is often wealthy). The Test variable refers to GPA and entrance exam results. The College variable stands for admissions into college and is highlighted in the conditional probability tables (CPTs) to indicate that it represents the decision being modeled. Job refers to whether the individual attained desirable employment. Note: For brevity, the CPTs for Test, College, and Job only contain the probabilities for when these values are one.

- Identify variables that morally justify unequal decisions (e.g., talent, ambition)
- Identify variables that are sensitive and do not justify unequal decisions (e.g., gender, socioeconomic status, race, etc).
- Identify variables over which you have control (e.g., college admissions).
- Identify the variable that represents a socially advantageous position (e.g., job)
- Identify variables in your data which you believe RAWLSNET should ignore.

To illustrate the method and purpose of RAWLSNET, we present a running example and discuss two uses of RAWLSNET: generating aspirational data and providing policy-advice to domain experts.

1.0.1 A running example. To illustrate FEO and RAWLSNET we use the following running example in the domain of education. We consider people applying for a job, where each applicant is born with a certain degree of talent and a certain socioeconomic status (SES). Subsequently, applicants achieve a certain secondary GPA and take college entrance exams. On the basis of these exam scores (including GPA) and their SES, applicants have a certain probability of getting into a prestigious college. After graduation, each applicant's college experience and SES determine how likely it is that he/she obtains a good job. We imagine that the initial situation, prior to RAWLSNET's intervention, reflects certain unjustified advantages for those with high SES. As is plausibly true for modern societies, high SES has an impact on test scores, college acceptance, and employment. Most relevantly, applicants with high SES have a better chance of achieving a good job than similarly talented people with low SES. This situation is a violation of FEO. In order to satisfy FEO, whether or not an applicant is given a job must be independent of the applicant's SES, given their talent. For exposition, we model all variables as binary and include only a single sensitive feature (which is SES), but note that neither of these limitations are necessary. Figure 1 shows the BN for our running example.

We consider the task of a hypothetical college admissions committee as they decide who is/isn't accepted into college based on the test scores and SES of potential applicants. Its goal is to achieve FEO, which we note is a consideration at the hiring level, not the college admittance level. By altering the probabilities of accepting low-SES applicants (based on their test scores), the college admissions committee can provide an advantage to low-SES applicants. This advantage can counteract the unfair benefits of high-SES students that directly influence both their test scores and job placement. RAWLSNET sets the probabilities $P(College|SES, Test)$ such that FEO is satisfied.

1.0.2 Aspirational data. RAWLSNET can be used to generate synthetic datasets that model ideally fair circumstances (a.k.a. *aspirational data*). Such data can be sampled from the FEO-satisfying BN models generated by RAWLSNET, and subsequently used to aid in research on algorithmic bias by shifting the focus towards sources of bias introduced downstream in the machine-learning life cycle, such as training and deployment processes. By evaluating data analysis methods on aspirational data, one can determine whether or not "bad data" is the only factor to blame in a biased system.

Complementary to our method of generating aspirational data are methods of debiasing or cleaning data [5, 7, 21]. Importantly,

however, RAWLSNET acts on the *distribution* from which the data comes, rather than the data itself. This yields two key advantages over debiasing methods. First, the resulting BN can be sampled from to generate aspirational datasets of any size and thus used to run a wider variety of experimental investigation than what any single debiased dataset could offer. Second, the alterations to the BN — which we note are changes to the conditional probabilities and not the structure of the BN — can be directly interpreted as policy advice, as we discuss next.

1.0.3 Policy advice. Another use of RAWLSNET is to aid in decision-making for policy makers. If decision-makers know the distribution of talent for their task, then RAWLSNET can be used to inform their decisions. Otherwise, the decision-makers may use RAWLSNET to evaluate how different policies will affect FEO under a variety of background assumptions regarding the distribution of talent. For instance, RAWLSNET might be used to advise acceptance decisions of a college admissions committee. The committee can evaluate the impact of different admissions policies in scenarios where talent is concentrated among few students, or where talent is equally distributed. This use of RAWLSNET is discussed in detail in Section 6.

1.0.4 Contributions. We introduce RAWLSNET, a method for altering a BN so as to satisfy Rawlsian FEO. To our knowledge, RAWLSNET is the first tool developed that produces aspirational (FEO) data distributions. RAWLSNET alters BN models in a way that preserves their initial structure. This preservation allows it to generate aspirational data for important social systems. Whenever satisfying FEO is not possible, RAWLSNET outputs a BN whose distribution is closest to satisfying FEO.

1.0.5 Paper structure. We provide background on FEO and its appropriate applications in the next section. Then, we discuss the formalization of our problem and present our contribution: RAWLSNET and its underlying mathematics. This is followed by experiments, related work, and discussion.

2 FAIR EQUALITY OF OPPORTUNITY

FEO requires that any two people with similar talent and ambition receive the same chance of achieving an advantageous social position (e.g., a good job), regardless of their background. This principle is designed to eliminate the effects of discrimination and other oppressive structures in determining who has access to advantageous social positions. It requires that social features irrelevant to determining who will do best at a job are irrelevant to determining who will receive the job.

FEO is one aspect of the Rawlsian theory *justice as fairness* [42, 44]. This theory is influential, well-supported, and widely popular [8]. FEO itself is particularly plausible, as similar principles appear in a variety of other theories [1]. Moreover, FEO has been used by philosophers to argue for social justice interventions including affirmative action [41, 49]. Note, also, that FEO is formally similar to other kinds of substantive equality of opportunity principles [22, 28, 45]. Thus, our work here can be applied to satisfying those principles as well.

Rawls [42, 43, 44] offers a theory of what is required for a society to be ideally fair. Rawls' theory of justice as fairness consists primarily of two principles. The first, which protects everyone's basic liberties (e.g., freedom of speech, religion, assembly), won't be our focus. The second principle governs the equitable distribution of wealth and social advantage. In particular, it concerns what inequalities are acceptable in an egalitarian society. The second principle itself consists of two parts: the first involves FEO, and the second is the *difference principle*. FEO governs who is eligible for unequal rewards, while the difference principle governs how unequal those rewards may be.

FEO requires that any wealth inequalities in society must be attached to advantageous social positions, what Rawls called *offices*. An office is desirable employment that carries with it greater responsibility, greater prestige, and/or higher pay. Rawls' theory requires that advantageous positions must be open to all applicants under FEO. Thus, FEO applies directly to decisions about employment. Obtaining an office is the only way, in a fair society, to become wealthier than your peers. In sum, FEO governs how good jobs are handed out.

Rawls called the sort of equality of opportunity we are interested in "fair" to contrast it with a more familiar sort, which is often called *formal equality of opportunity* (or formal EO for short). Formal EO requires two things: (1) that positions of social advantage be open to all applicants, and (2) that applicants be evaluated entirely based on their qualifications for the position [2, 44]. Formal EO rules out explicit discrimination based on group membership (e.g., race and gender). It also rules out caste systems, nepotism, and favoritism. Formal EO is essentially an ideal version of what contemporary equality of opportunity laws governing housing and employment aim at. While itself quite stringent, formal EO is compatible with a wide variety of oppressive structures and implicit discrimination. For instance, the hiring process for an engineering position might conform to formal EO if it widely publicizes its openings, considers all applications, and hires the most qualified engineers. But if only men are allowed to attend engineering schools, the result is still unfair. Formal EO is an important principle of fairness. RAWLSNET is designed to recommend policies that will help satisfy FEO without creating violations of formal EO. It achieves this by suggesting alternative policies for decisions (such as college admissions) that occur prior to handing out advantageous positions (such as hiring decisions). Thus, formal EO for hiring decisions is maintained.

The primary difference between formal EO and FEO concerns which features of an applicant are relevant to justifying hiring choices. Formal EO focuses on qualifications: the skills, training, and experience that an applicant has *at the time of hiring*, which determine how good the applicant would be at the job. In a real-world, contemporary society, sensitive but morally irrelevant features can make a significant difference to what qualifications an applicant has. Being born into a high SES family, for instance, has a large impact on the kind of education one has access to. In the imagined (but realistic) engineering case mentioned above, it was gender that was inappropriately affecting educational opportunities. It is the focus on qualifications that makes formal EO inadequate to fully ensure genuine fairness. Yet qualifications are clearly enormously important: we do not want our bridges built by engineers who did not go to engineering school.

FEO is a principle meant to fill the gap between what formal EO requires and what is required for a genuinely fair EO (hence the name). It focuses not on qualifications at the time of hire, but

instead on innate *talent* and *ambition*. Here, we can understand talent as an innate potential to be good at some job. Ambition is one's willingness to develop and use their talent. For most of us, no amount of training, experience, and hard work could make us into basketball players as good as Lebron James. What he has, in addition to an incredible work ethic and ambition, is innate talent most of us lack. Similarly, there are many other innate features which effect an individual's potential to excel at various jobs. Following Rawls, we label these features, generically, as *talent*. A person's willingness to work to develop and employ his/her talents constitutes his/her ambition. For brevity, we will generally use one variable, labeled "talent," to represent an individual's innate talent and ambition.

According to Rawls, FEO requires that "those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the system" [43, p. 63]. Two people with the same talent and degree of ambition should have equal chances of obtaining desirable employment, and the social benefits it brings with it. In other words, one's chance of getting a good job should be statistically independent of any features of an individual other than his/her talent and ambition.[1] FEO requires that features such as ethnicity, gender, LGBTQ+ status, nationality, birthplace, etc., must all be statistically irrelevant to whether one achieves an office. We call these *sensitive features*. Only talent and ambition should ultimately make a difference to whether you get the job.

Despite the focus on talent and ambition, Rawls' theory of fairness is not a meritocratic theory. That is, Rawls does not suggest that having greater innate talent makes a person more deserving of greater rewards. Innate talent is not something one earns or deserves credit for, and is therefore "arbitrary from a moral point of view" [42, p. 72]. However, it can benefit everyone in society for talented individuals to develop their talents, assuming the increased wealth produced by deploying such talents is distributed equitably. (This equitable distribution is ensured by the other part of Rawls' second principle, the difference principle). The justification for the appeal to talent in FEO, then, is not that talented people deserve more wealth. Instead, the justification concerns improving things for everyone. FEO is a principle that is tailored to eliminate unfair advantages that occur as a result of social circumstances, while allowing that talented individuals may be be incentivized to develop and use their talents for everyone's benefit.

FEO is perfectly compatible with thinking that there are no innate differences in talent between individuals. In that case, FEO will simply require that all individuals have an equal chance at achieving an advantageous social position. RAWLSNET can easily handle this assumption as well. Thus, we are not committed to thinking there are innate talent differences. However, FEO is compatible with the assumption that there are. There is some empirical evidence supporting the existence of innate talent differences from the psychology of expertise [19].

An extremely plausible (and perhaps morally obligatory) assumption is that innate talent is distributed independently of sensitive features. Friedler et al. [2016] call this the *We are All Equal* assumption. There is no significant or compelling evidence to think innate talent differences track sensitive categories such as race. Meanwhile, there is strong evidence that various achievement gaps can be explained in purely environmental terms [35–37]. There are thus compelling moral and epistemic reasons to assume there are no such inter-group differences [45]. This assumption is commonly made in works that formalize substantive equality of opportunity principles such as FEO [16, 22, 29, 30].

According to the *We are All Equal* assumption, any differences in job candidates' qualifications that are correlated with sensitive features must be the result of differences in experience and training. FEO requires that such differences be made ultimately irrelevant to determining who receives an advantageous position. Any two applicants with the same talent and ambition should have the same probability of obtaining desirable employment. Satisfying FEO therefore requires removing or ameliorating the impact of sensitive features on employment.

As noted above, FEO must be satisfied in a way that avoids violating *formal* EO. Satisfying both principles requires changing the way talent and ambition are related to qualifications.[2] RAWLSNET is designed to allow a decision-maker to satisfy FEO (when possible) without creating violations of formal EO. It accomplishes this because it is designed to operate on decisions that are made *prior to the point of hiring for advantageous positions*. Such decisions are not directly governed by formal EO, as that principle concerns employment decisions exclusively. Our strategy is illustrated by our college admissions example. There, the decision in question is whether to admit someone to a prestigious college. But being a student at such an institution is not an office: it is not a job by which unequal wealth is distributed. Admissions do indirectly impact social advantage, but only insofar as they impact hiring.

3 PROPOSED APPROACH: RAWLSNET

In this section we define what it is to be an *FEO Application*. We also provide guidance for determining whether a task is an FEO application. We then present RAWLSNET, a method for altering the parameters of a BN model for an FEO application in order to satisfy FEO. We also discuss runtime considerations and encoding constraints of the underlying application.

3.1 FEO Applications

FEO governs the distribution of advantageous social positions. Applied to our running example involving college admissions, RAWLSNET is designed to determine the correct college acceptance rates (i.e., the probability of being admitted) to ensure that FEO is satisfied at the later stage of hiring. Our running example is what we call an *FEO application*. "FEO application" is a novel term which we define as: *a decision which affects whether FEO is satisfied by a distinct, subsequent hiring decision*. Genuine FEO applications are decisions that are needed to satisfy FEO, but which do not introduce violations of formal equality of opportunity. We avoid violations of formal EO by using earlier decisions to improve the qualifications of talented applicants prior to the hiring process. In the college

[1] Strictly speaking, what is required is that any two applicants with the same talent and ambition should have the same probability of getting the job *when not conditioned on other attributes such as training*.

[2] There is significant dispute about the relationship between formal and fair equality of opportunity. In particular, there is dispute about whether satisfying FEO requires also satisfying formal EO [2, 51, 52]. We seek to sidestep this dispute. We take both principles to be important, and remain neutral on their relative priority.

admissions example, this involves making it more likely that talented students from low SES backgrounds are admitted to college, which improves their qualifications for being hired. Thus, in our running example we satisfy FEO without violating formal EO by intervening on the earlier admissions decision.

3.1.1 A Guide for Recognizing FEO Applications. For a decision to count as an FEO application, it must meet three conditions: it must (1) affect the distribution of advantageous social positions (i.e., good jobs), (2) be made prior to hiring decisions, and (3) be made on the basis of appropriate features of applicants. This is illustrated by our college admissions example, which meets all of these conditions. These are *necessary* conditions for being an FEO application. Moreover, any decision that meets these conditions is likely to be such an application. Regarding condition (1), recall that offices are desirable employment positions that carry with them the social advantages that lead to inequality. Condition (2) helps ensure that the decision-maker does not introduce violations of *formal* EO. Condition (3) concerns the relevant features of applicants. In an FEO application, we distinguish three categories of relevant features. There are *justifying* features: features that morally justify the inequalities which are attached to advantageous positions. Following Rawls, these are limited to talent(s) and ambition. There are also *sensitive* features: features which should not be allowed to lead to inequalities. As mentioned, these are features such as race, gender, and socioeconomic status. Finally, there are what we simply call "other" features. Other features may permissibly impact hiring and other decisions, but which do not themselves justify inequality. For the purposes of recognizing an FEO application, its crucial to identify the relevant justifying features and sensitive features.

In our college admissions example, the *other* category includes the exam score feature and the college admissions decision itself. In fact, this is an important feature of an FEO application decision: it will always concern a feature that is neither sensitive nor justifying. This is because, as we have defined the term, an FEO application is always a decision by someone other than the applicant, and is one that indirectly impacts the distribution of advantageous positions.

3.2 Altering Bayesian Networks with RAWLSNET

RAWLSNET provides a method for altering a given BN (describing an FEO application) such that its defined distribution satisfies the Rawlsian FEO principle. RAWLSNET can work when data is provided with or without a BN. In the latter case, it will learn a BN structure that best fits the data. In the discussion to follow, we assume a BN structure has been given along with the data with variables $\mathcal{V}$, edges $\mathcal{E}$, and parameters $\mathcal{P}$ – i.e., elements of the conditional probability tables (CPTs) of each variable. The variables are partitioned as $\mathcal{V} = \boldsymbol{J} \cup \boldsymbol{S} \cup \boldsymbol{O}$ where $\boldsymbol{J}$ is the set of variables morally justified for inequality (e.g., talent), $\boldsymbol{S}$ is the set of sensitive variables (e.g., socioeconomic status), and $\boldsymbol{O}$ is the set of remaining (other) variables. In addition, a *control* variable $C \in \boldsymbol{O}$ is specified as the variable which we can control the CPT of (e.g., college admissions) and a *target* variable $Q \in \boldsymbol{O}$ (e.g., obtaining a good job). All variables must be categorical because we model discrete probability distributions. Further, we assume that the justified and sensitive variables, given their innate nature, are root nodes in the BN structure. This assumption ensures that we solve a linear system of equations. One can optionally supply RAWLSNET a set of *feasibility constraints*, which we discuss below. Table 1 summarizes our notation.

In this context, FEO is obtained by ensuring statistical independence between the target variable Q, and the sensitive variables $\mathbf{S}$, conditioned on the justified variables $\mathbf{J}$. That is, we seek to set the CPT values of the control variable C such that:

$$Q \perp\!\!\!\perp \boldsymbol{S} \mid \boldsymbol{J} \tag{1}$$

Equivalently, we seek values for the conditional probability of the control variable C given its parents such that:

$$P(Q \mid \boldsymbol{j}) = P(Q \mid \boldsymbol{j} \cup \boldsymbol{s}) \tag{2}$$

for every possible assignment $\boldsymbol{j}$ of variables in $\boldsymbol{J}$ and assignment $\boldsymbol{s}$ of variables in $\boldsymbol{S}$. For any particular assignments $\boldsymbol{j}, \boldsymbol{s}, \boldsymbol{o}$ of justified, sensitive, and other variables respectively, we let:

$$\mathsf{Par}(V; \boldsymbol{j}, \boldsymbol{s}, \boldsymbol{o}) \tag{3}$$

denote the set of variables (and their assignments in $\boldsymbol{j}, \boldsymbol{s}, \boldsymbol{o}$) that are the parents of V.

Note that (2) is a collection of desired qualities. For ease of notation, we focus on just one. Namely, our goal is to satisfy:

$$P(q \mid \boldsymbol{j}) = P(q \mid \boldsymbol{j}, \boldsymbol{s}) \tag{4}$$

We first note that:

$$P(q \mid \boldsymbol{j}) = \alpha P(q, \boldsymbol{j}) \tag{5}$$

$$= \alpha \sum_{\tilde{\boldsymbol{o}}} \sum_{\tilde{\boldsymbol{s}}} P(q, \boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}}) \tag{6}$$

where α is a normalization constant, and $\tilde{\boldsymbol{o}}$ ranges over all possible assignments for $\boldsymbol{O}$ (resp. $\tilde{\boldsymbol{s}}, \boldsymbol{S}$). We let $Z^{(q)}_{\boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}}} = P(q \mid \mathsf{Par}(Q; \boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}}))$ and for any set $\boldsymbol{V}$ of variables and assignment $\boldsymbol{a}$ of its parents we let:

$$Z^{(\boldsymbol{V})}_{\boldsymbol{a}} = \prod_i P(V_i \mid \mathsf{Par}(V_i; \boldsymbol{a})) \tag{7}$$

For ease of notation, we suppress the dependency on the assignment when clear and simply write $Z^{(q)}, Z^{(\boldsymbol{S})}$. With this notation we have:

$$P(q \mid \boldsymbol{j}) = \alpha \sum_{\tilde{\boldsymbol{o}}} \sum_{\tilde{\boldsymbol{s}}} Z^{(q)} Z^{(J)} Z^{(S)} Z^{(O)} \tag{8}$$

Separating out the value in $Z^{(O)}$ corresponding to C we let:

$$Z^{(O)}_{\boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}}} = Z^{(O')}_{\boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}}} P(c | \mathsf{Par}(C; \boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}})) \tag{9}$$

to obtain:

$$P(q \mid \boldsymbol{j}) = \alpha \sum_{\tilde{\boldsymbol{o}}} \sum_{\tilde{\boldsymbol{s}}} Z^{(q)} Z^{(J)} Z^{(S)} Z^{(O')} P(c | \mathsf{Par}(C; \boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}})) \tag{10}$$

Similarly, for the right-hand side of (4) we have:

$$P(q \mid \boldsymbol{j}, \boldsymbol{s}) = \tilde{\alpha} \sum_{\tilde{\boldsymbol{o}}} Z^{(q)} Z^{(J)} Z^{(S)} Z^{(O')} P(c | \mathsf{Par}(C; \boldsymbol{j}, \tilde{\boldsymbol{o}}, \tilde{\boldsymbol{s}})) \tag{11}$$

Therefore we can satisfy Rawlsian FEO by satisfying the equality (10) = (11) for each assignments $\boldsymbol{j}$ and $\boldsymbol{s}$. This yields a system of linear equations, solvable using standard methods. Equations for which the normalization constants are undefined due to division by zero, as is the case when talent is equal for all individuals, are omitted from the system.

Symbol	Meaning
$\boldsymbol{J} = J_1, \ldots, J_{\lvert\boldsymbol{J}\rvert}$	The set of justified variables
$\boldsymbol{S} = S_1, \ldots, S_{\lvert\boldsymbol{S}\rvert}$	The set of sensitive variables
$\boldsymbol{O} = O_1, \ldots, O_{\lvert\boldsymbol{O}\rvert}$	The set of other variables
$\boldsymbol{v} = v_1, \ldots, v_{\lvert\boldsymbol{V}\rvert}$	An assignment to the variables in the corresponding set $\boldsymbol{V}$
$\texttt{Par}\ (V; \boldsymbol{j}, \boldsymbol{s}, \boldsymbol{o})$	The set of variables that are parents of V, and their assignments in $\boldsymbol{j}, \boldsymbol{s}, \boldsymbol{o}$
Q, q	The target variable, a particular assignment of it
C, c	The control variable, a particular assignment of it

Table 1: Notation table. RAWLSNET is provided a BN with variables $\boldsymbol{J} \cup \boldsymbol{S} \cup \boldsymbol{O}$. In addition, a control variable $C \in \boldsymbol{O}$ and target variable $Q \in \boldsymbol{O}$ are specified. The output of RAWLSNET is a new BN, identical to the original in structure and all parameters except select elements of the CPT for C, which are edited such that Rawlsian FEO is satisfied for the target variable Q.

3.2.1 Feasibility Constraints. Consider our running example of college admissions. Using RAWLSNET, one can select an admissions policy so as to ensure fair (in the Rawlsian sense) job allocation. However, if we solve the system of equations described above, there is no guarantee that the resulting admission policy is satisifiable in practice — it may require that the school admits substantially more (or fewer) students than is viable. We therefore imbue RAWLSNET with the capability to accept a set of constraints.

Given an equality constraint on some marginal distribution of the BN (e.g., the expected number of students admitted to college is precisely p percent of the population), we simply add this to the collection of equations defined above. With similar derivation to (5) through (10), we note that the above constraint is also linear and thus the system is still efficiently solvable. Given inequality constraints (e.g., a particular marginal probability must be in some interval), RAWLSNET solves a linear program.

3.2.2 Runtime. We note the following runtime considerations. At its core, RAWLSNET is solving the linear system above. The two bottlenecks in terms of runtime are (a) the number of constraints and (b) the time it takes to construct the constraints. For every assignment of $\boldsymbol{J}$ and $\boldsymbol{S}$, we have a constraint of the form in Equation (4). While this is exponential in $|\boldsymbol{J}| + |\boldsymbol{S}|$, we note that in practice both sets are typically small. To compute the coefficients (i.e., the $Z^{(\cdot)}$ terms in (10) and (11)), we perform exact inference in the underlying BN. This is doable in linear time via dynamic programming if the BN is a polytree [48]. We note that approximate inference (e.g., via particle filtering) is an option for general Bayesian Networks.

4 EXPERIMENTS

We demonstrate the effectiveness of RAWLSNET on the following data: (1) our illustrative college-admissions example (2) another illustrative example regarding campaign financing for elections (3) a synthetic HR dataset examining employee promotions by IBM, and (4) a real campus recruitment dataset . All experiments were run on a laptop with 8GB memory and a 1.8 GHz Intel Core i5 processor. We utilized cvxpy to optimize constrained linear systems of equations, pgmpy for Bayesian Network training and inferences, and matplotlib for plotting. The code is available at https://github.com/dliu18/rawlsnet.

4.1 Illustrative Example: College Admissions

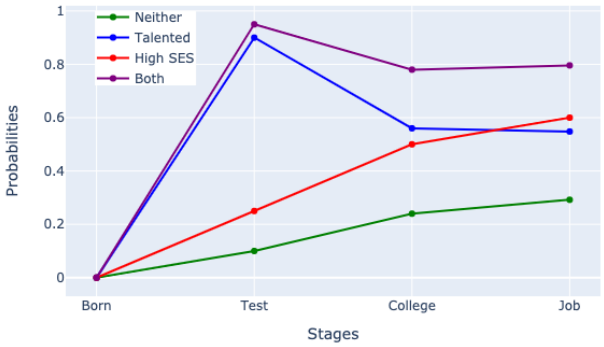
We start with our college admissions example, where the BN is specified in Figure 1. We assign variable names to each of the nodes in the BN. T is Talent, SES is Socioeconomic Status, E is Test (Exam) Score, C is College, and J is Job. With this notation, the justified, sensitive, and other variables are as follows: **J**= {T}; **S**= {SES}; **O**= {E, C, J}. The control variable is the college admissions policy C and the query variable is job J. To satisfy FEO, whether someone gets a job must be independent of their SES given their Talent. Therefore, the following equalities must hold: $P(J|T$=Low, SES=Low) $\equiv P(J|T$=Low, SES=High) and $P(J|T$=High, SES=Low) $\equiv P(J|T$=High, SES=High). Substituting Equation 11 into these equations yields a linear system of equations, where the variables are the CPT values for the College node. RAWLSNET solves this linear system.

Results. Figure 2 visualizes the distribution of those receiving good jobs before and after we perform RAWLSNET. As the figure shows, both distributions share the same testing and hiring policies that have been chosen to encode well-known societal biases, such as the fact that those with high SES backgrounds generally fare better in the hiring process than those of low socioeconomic background.

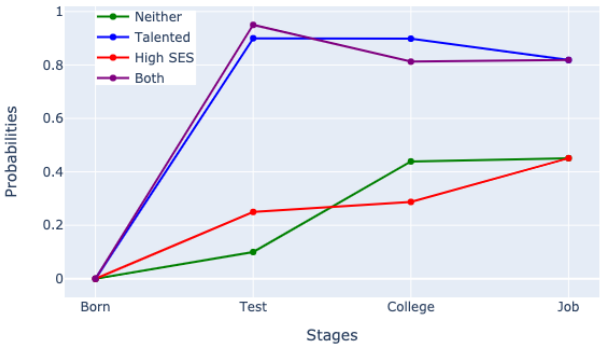
We highlight the following flexibility of RAWLSNET. Two of the four CPT values for college admission are fixed, and RAWLSNET is tasked with determining the other two values. As a result, the solution for this particular example can be interpreted as: given the admissions policy for high SES individuals, what policy must be implemented for the low SES applicants to ensure FEO at the job application phase?

The probabilities of college-admissions and job-offers in Figure 2(a) agree with our expectations of an unfair world. However, the probabilities after RAWLSNEThas been applied in Figure 2(b) fixes this unfairness – as indicated by the intersection of the blue and purple lines as well as the red and green lines.

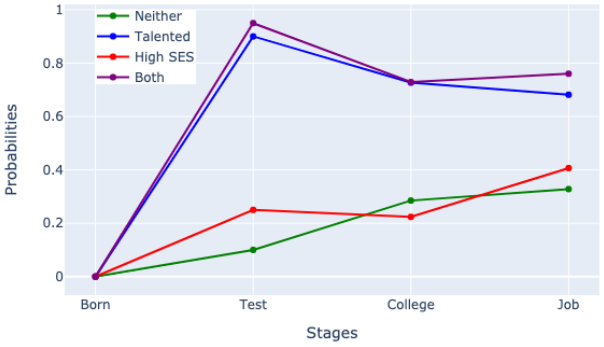
We highlight an additional capability of RAWLSNET: feasibility constraints. In Figure 2(c), we show the output of RAWLSNET when the aggregate college-admissions rate is capped at 50%. When such (linear) constraints are at play, RAWLSNET solves a linear program. As expected, the conditional probabilities for Job are no longer equal (as FEO is not achievable under these constraints), but they are much closer to FEO compared to the original distribution (as seen in Figure 2(a)). We note that RAWLSNET obtaining probabilities as close as possible to an FEO satisfying distribution is especially helpful when it is used to inform policy makers. In this example,



(a) Probability of desired outcome before applying RAWLSNET.



(b) Probability of desired outcome after applying RAWLSNET.



(c) Probability of desired outcome after applying RAWLSNET with feasibility constraints.

Figure 2: Visualizations of the outcomes of the synthetic college-admissions BN. The x-axis lists each successive desired outcome and the y-axis represents the proportion of the population that obtained the desired outcome. The population is further broken down on the basis of SES and talent. Except for the college-admissions CPT, the other CPTs were identical. RAWLSNET can achieve FEO with the probability of obtaining a job being independent of SES.

the college-admissions committee can better understand how the cap on the admissions rate affects their ability to satisfy FEO.

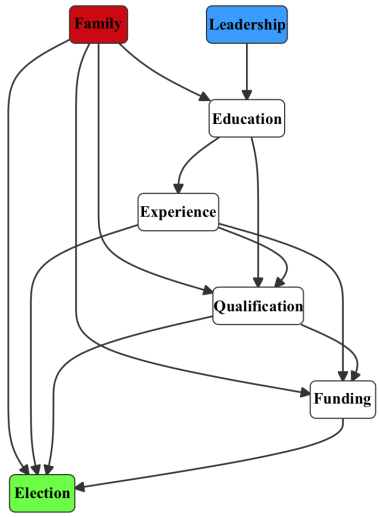


Figure 3: The synthetic campaign finance example sets the probability of being born with *Leadership* skills as the *justified* variable. The probability of belonging to a *Family* with political influence is chosen as the *sensitive* variable. *Funding* is the *control* variable. *Election* is the advantageous social position (i.e., the *target* variable). The variables *Education*, *Experience*, *Qualification* are *other* variables.

4.2 Illustrative Example: Campaign Finance

To highlight a situation where strict FEO is not satisfiable, we consider the domain of financing a political campaign and show the BN structure for this example in Figure 3. We choose the values for the CPTs based on our understanding of the real-world dynamics of an election campaign, where we assume that those who come from a family of politicians, have a better chance of obtaining funding and winning elections than those who do not. All variables in this example are binary, except the target variable of *Election* which can take on three values: "Not Elected", "Nominee", and "Elected".

Results. Table 2 shows the original probabilities for the *Election* variable given *Leadership* and *Family*. The goal of RAWLSNET is to modify the CPT for the *Funding* variable such that for a given *Leadership* level (i.e., "Good" or "Poor"), the probability of being elected or nominated does not depend on whether or not a person comes from a family with political history. Given the initial CPTs, RAWLSNET manages to reduce the disparity between the probabilities of winning given leadership and family background (see Table 3). However, the conditional probabilities do not satisfy Equation 4. To solve the system of equations, seven of the eight CPT values for the *Funding* variable would need to be greater than one. Instead, RAWLSNET provides valid CPT values that come closest to satisfying Equation 4, where closest refers to the CPT values that minimize the squared-difference between Equations 10 and 11.

4.3 Synthetic Data: IBM HR Dataset

The IBM HR Analytics Employee Attrition & Performance dataset [50] is a synthetic dataset created by IBM to model the factors that lead to employee attrition. We use the dataset to model gender-bias in staff

Lead.	Family	Not Elected	Nominee	Elected
Poor	Not Political	29.20%	33.20%	37.60%
Poor	Political	19.1%	8.60%	72.30%
Good	Not Political	27.5%	30.20%	42.30%
Good	Political	17.8%	8.10%	74.10%

Table 2: Original probability values for $P(Election|Leadership, Family)$ for the Campaign Finance example.

Lead.	Family	Not Elected	Nominee	Elected
Poor	Not Political	16.34%	24.10%	59.50%
Poor	Political	22.34%	11.10%	66.60%
Good	Not Political	15.96%	21.40%	62.60%
Good	Political	21.14%	10.70%	68.20%

Table 3: Updated probability values for $P(Election|Leadership, Family)$ after using RAWLSNET for the Campaign Finance example. Note that FEO cannot be satisfied in this case, thus RAWLSNET selects the closest possible distribution.

promotions. *Gender* is the *sensitive* variable. *Education* is a proxy for talent; thus it is the *justified* variable. *RecentPromotion* is the advantageous social position (i.e., the *target* variable). We assume that work-life balance (named $WorkLifeBalance$) and $JobSatisfaction$ are the *other* variables, with the $WorkLifeBalance$ also being the *control* variable. The relationships between the variables are shown in Figure 4. This data set contains 35 features out of which 4 are related to the FEO task at hand. The variables *Education* and $JobSatisfaction$ are categorical with 5 and 4 categories, respectively. Higher values correspond to greater education and satisfaction. The variable *RecentPromotion* refers to the number of years since the employee was last promoted. We convert *RecentPromotion* into a binary variable by thresholding at the median value. The variable $WorkLifeBalance$ is also a categorical variable with 4 categories. We convert it into a binary variable for ease of exposition.

Results. Table 4 shows the original probabilities of getting a promotion given an education level (Bachelors, Masters, etc.) and gender. We observe slight discrepancies between the promotion probabilities of male and female employees. Table 5 shows the results after RAWLSNET was applied to this network, which updated the values of the CPT for the $WorkLifeBalance$ node. The results show that given an education level, the probability of promotions stay the same regardless of gender.

4.4 Real Data: Campus Recruitment

The Campus Recruitment dataset from Kaggle [47] contains information about students in India. It includes students' scores from standardized testing, whether or not they got a job at the end of school, and with what salary. While the data set is not primarily designed for making decisions which count as FEO applications, it includes relevant data and is publicly available. It involves at least one decision that affects the distribution of advantageous social

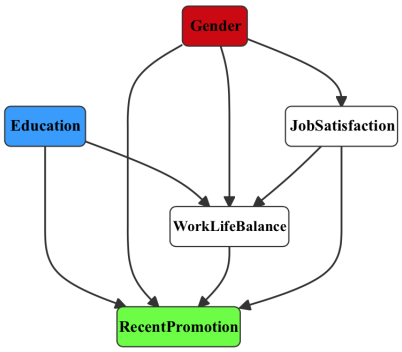


Figure 4: IBM HR Example: We consider the node *Education* as a proxy for talent (thus, it is the *justified* variable). *Gender* is the *sensitive* variable. $WorkLifeBalance$ is the *control* variable since we assume that employers can alter it. *RecentPromotion* is the advantageous social position within the company, and hence is the *target* variable. $JobSatisfaction$ is an *other* variable.

Education	Gender	Promotion
Below College	Male	36.27%
Below College	Female	37.17%
College	Male	36.27%
College	Female	37.17%
Bachelor	Male	41.79%
Bachelor	Female	39.87%
Master	Male	40.19%
Master	Female	40.73%
Doctor	Male	40.19%
Doctor	Female	36.51%

Table 4: Original probability values for $P(RecentPromotion|Education, Gender)$ for the IBM HR data.

positions: whether a student receives a competitive internship. We assume that students receiving internships have a better chance at landing a job later. Thus, we use this decision as an example of an FEO application. Note, however, that the *justified* variable, *SchoolPercent*, is an imperfect proxy for talent. It represents the earliest standardized test score available for each student. We discuss important caveats for using RAWLSNET in cases like this in Section 6.

We assume *Gender* is the *sensitive* variable. *Internship* refers to whether the student received a competitive internship and is our *control* variable. *Salary* is the advantageous social position (i.e., the *target* variable). The BN shown in Figure 5 also includes variables *DegreePercent* that represents the undergraduate scores, *HighSchoolPercent* that represents standardized test scores during high school, and *EmploymentTest* that stands for scores earned in a

Education	Gender	Promotion
Below College	Male	32.61%
Below College	Female	32.61%
College	Male	32.61%
College	Female	32.61%
Bachelor	Male	43.04%
Bachelor	Female	43.04%
Master	Male	38.33%
Master	Female	38.33%
Doctor	Male	34.80%
Doctor	Female	34.80%

Table 5: Updated probability values for $P(RecentPromotion|Education, Gender)$ after using RAWLSNET for IBM HR data.

test that determines eligibility for the job. These variables constitute the *other* variables.[3]

The datatset consists of 15 features of which we use 7 that are relevant to the FEO-use case. The variables *SchoolPercent*, *DegreePercent*, *HighSchoolPercent*, *EmploymentTest* and *Salary* are continuous variables. We used the median values for each of these variables to convert them to binary variables. The variables *Gender* and *Internship* are categorical variables with a cardinality of 2 that contain strings, which were converted to numeric categorical variables. The relationships among these variables are shown in Figure 5. The CPTs for each node were learned from the data using maximum likelihood estimation.

Results. Table 6 shows the original probabilities of getting a good job given *SchoolPercent* (i.e., the talent proxy) and *Gender*. The table shows that male applicants have a higher probability of getting a good salary as compared to female applicants given the same talents. Table 7 shows the probabilities of getting a job given talent and gender after RAWLSNET has modified the CPT for the control variable *Internship*. We observe that FEO is satisfied and the probabilities of getting a good salary remain the same irrespective of gender.

SchoolPercent	Gender	Salary
Low Score	Male	50.86%
Low Score	Female	30.59%
High Score	Male	61.87%
High Score	Female	33.89%

Table 6: Original probability values for $P(Salary|SchoolPercent, Gender)$ for the Campus Recruitment Data.

5 RELATED WORK

In recent years, Rawls' work has become influential in the algorithmic fairness literature [3, 31, 34]. Some of this work focuses

[3]The variable names have been changed to improve readability. They do not necessarily match the ones in the original dataset.

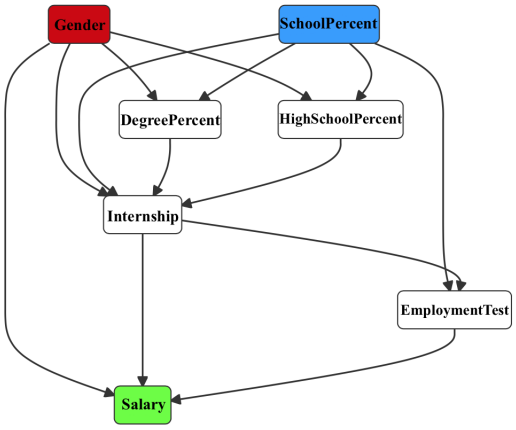


Figure 5: The campus recruitment data looks at the probability of getting a job with a good salary given the innate talent and gender of the applicant. We consider *Gender* as the *sensitive* variable and *SchoolPercent* as the *justified* variable. The nodes *HighSchoolPercent*, *DegreePercent*, and *EmploymentTest* are test scores for high school, undergraduate degree and an employment eligibility test, respectively. They all belong to the *other* variables. *Internship* looks at distributing internships to students which could help their prospects for a job at a later time. This is the *control* variable. *Salary* is the advantageous social position and the *target* variable.

SchoolPercent	Gender	Salary
Low Score	Male	43.06%
Low Score	Female	43.06%
High Score	Male	47.76%
High Score	Female	47.76%

Table 7: Updated probability values for $P(Salary|SchoolPercent, Gender)$ after using RAWLSNET for the Campus Recruitment Data.

on using the other aspects of Rawls' theory, such as the original position [42], to develop novel principles of governance to ensure appropriate transparency, explainability, and fairness [18, 27, 53]. Other work has appealed to Rawls' difference principle or prioritarian principles inspired by it [11, 13, 26]. These works share a general philosophical outlook with our project, as they concern justice as fairness. However, they appeal to different parts of the Rawlsian framework to achieve different goals. Our work is complementary to these others as it implements another aspect of Rawls' theory. Together, these approaches offer the potential for a unified contractualist approach to the ethics of AI. In contrast, Lundgard [31] raises objections to use of Rawls' theory for fair ML, but these objections are less pressing for FEO in particular.

There has also been significant interest in complementary projects in the algorithmic fairness literature which appeal to substantive equality of opportunity principles, including FEO [14, 16, 20, 23, 24, 33, 55]. These appeals are used to justify various specific fairness

metrics, and to adjudicate disputes between these metrics. Work in this literature appeals to Rawls' FEO and similar substantive equality of opportunity principles, in particular those discussed and formalized by John Roemer [45, 46]. The primary difference between our project and these others is that they are concerned with determining appropriate metrics of fairness. In other words, they are looking to measure the degree to which particular uses of ML algorithms count as fair. They use these metrics to evaluate and mitigate bias in ML. Many of those who appeal to FEO do so in order to argue for one or another fairness metric as better than alternatives, or at least better for a particular type of circumstance. For example, Binns argues that FEO considerations justify appeals to group fairness metrics in certain contexts and individual fairness metrics in other contexts [6]. Loi et al. argue that a modified, generalized version of FEO justifies two different fairness metrics, sufficiency and separability, in distinct contexts [30]. Heidari et al. similarly argue that various notions of algorithmic fairness can be justified as special cases of a substantive EO principles like FEO [22].

Our project offers a useful addition to these other substantive equality of opportunity approaches. RAWLSNET is novel in that it models interventions on decisions. It computes which interventions will obtain FEO (or promote it as much as possible). Thus, the goal of our project is significantly different than the goal of those in the fairness-metric literature. In addition, RAWLSNET's output distribution can be sampled to generate new aspirational data, or it can be used to inform decision-makers. It is less concerned with evaluating the performance of ML algorithms, though it might be useful for that purpose, as we hope to explore in future work.

Previous efforts in training fair BNs achieve fairness by either retraining the BN with re-labeled data [32] or imposing fairness constraints during parameter learning [12]. RAWLSNET instead directly modifies the appropriate conditional probability values without changing the structure of the BN. Unlike past approaches [10], our goal is to generate fair data distributions, which can subsequently be used for sampling aspirational data or guiding policy decisions.

Previous attempts have also been made to generate fair data with other models, such as GANs [54]. However, our approach is unique in basing our definition of fairness on Rawls and in producing aspirational data distributions. As such, we are able to provide clearer guidance on when aspirational data generated through RAWLSNET should and should not be used.

6 DISCUSSION

We presented RAWLSNET: a method that determines how a BN model of an FEO application must be altered in order to satisfy FEO. RAWLSNET offers the ability to model circumstances of ideal fairness in order to generate distributions over aspirational (FEO) data. This aspirational data distribution can be used by researchers to promote fair ML by sampling from it to discover and avoid pitfalls in ML algorithms, which can lead to unfairness despite unbiased data.

RAWLSNET can also be used to offer advice to decision-makers seeking to promote FEO. However, caution must be exercised when using RAWLSNET for this purpose. In most circumstances, RAWLSNET should only be used for indirect advice: evaluating courses of action under a variety of hypothetical circumstances. The system's direct advice will reliably promote fairness only if practitioners have a reliable, unbiased proxy for talent. In most cases this condition will not be satisfied, given the empirical difficulties in isolating the impact of innate talent on measures of achievement such as standardized testing.

The accuracy of the results of RAWLSNET depend on the accuracy of the assignments of the variables in the data to the appropriate justified, sensitive, and other categories. Talent refers to an individual's innate, intrinsic features which partially determine their capability for succeeding and excelling in a social position. These features are only indirectly observable. Moreover, the actual evidence we have regarding innate talent will often be confounded by the complexity of the social systems which impact individual education and development. For instance, a good proxy for talent may be early standardized testing. However, a student's test scores will also certainly be influenced by their early home life, which is in turn influenced by factors like socioeconomic status. Thus, the very bias we are attempting to eliminate may creep into the proxy we use to evaluate talent.

Our use of the campus recruitment data [47] illustrates some of the problems for using RAWLSNET for policy advice. In order to make use of a public data set we used an imperfect proxy for talent. We assigned the "SchoolPercent" feature, which refers to standardized test scores, to the justified variable category on the assumption that it was the best available proxy for talent. However, in this case the proxy will certainly be imperfect, as sensitive features are likely to impact this test score. RAWLSNET would do a better job at advising policy decisions to promote FEO with a better proxy for talent. As it stands, the recommendations of RAWLSNET for this example may help to promote FEO, but they will not be able to guarantee FEO is satisfied without a better, less biased proxy for talent. In future research, we intend to pair RAWLSNET with methods that infer the distribution of unobserved features such as innate talent. In the meantime, taking RAWLSNET's output as direct advice should be avoided. RAWLSNET can be safely used to provide indirect advice for policy makers and domain experts. It can provide answers to hypothetical questions regarding how a policy will fare with respect to FEO, assuming different possible distributions of talent in the relevant population. To illustrate this, imagine a group of political scientists developing policy proposals for ameliorating the effect of SES on standardized test scores. They have done a number of empirical studies to gather data in different parts of the country. Using a variety of ML methods, they create a set of Bayesian Networks that model different possible ways the world could be, consistent with their data. They are not sure which one is the most accurate. These scientists could use RAWLSNET to find the best policy to promote FEO in each one of these BN models. If the one policy for giving out extra training is successful in promoting FEO in all their BN models, this could provide support for implementing that policy. Alternatively, it might be that a policy works well in some circumstances but fails disastrously in others. This would then be a sign that more research is needed before implementing the policy. In this way, RAWLSNET provides useful information, despite the fact that one cannot merely take its output as a proposed policy.

REFERENCES

[1] Richard Arneson. 2015. Equality of Opportunity. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Stanford University.

[2] Richard J. Arneson. 1999. Against Rawlsian Equality of Opportunity. *Philosophical Studies* 93, 1 (1999), 77–112.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

[4] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017).

[6] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *FAT**. 514–524.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS*. 4349–4357.

[8] Harry Brighouse. 2005. *Justice*. Polity.

[9] Meredith Broussard. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. The MIT Press.

[10] Rodrigo L. Cardoso, Wagner Meira Jr., Virgílio A. F. Almeida, and Mohammed J. Zaki. 2019. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *AIES*. 437–444.

[11] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *FAT**. 339–348.

[12] YooJung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy Van den Broeck. 2020. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns. In *AAAI*. 10077–10084.

[13] Neelke Doorn. 2010. A Rawlsian Approach to Distribute Responsibilities in Networks. *Science and Engineering Ethics* 16, 2 (June 2010), 221–249.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proc. of the 3rd Innovations in Theoretical Comp. Sci. Conf.* 214–226.

[15] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016).

[17] Hannah Fry. 2018. *Hello World: Being Human in the Age of Algorithms*. W. W. Norton & Company.

[18] Jamie Grace. 2020. 'AI Theory of Justice': Using Rawlsian Approaches to Better Legislate on Machine Learning in Government. *Available at SSRN 3588256* (2020).

[19] David Z Hambrick, Brooke N Macnamara, Guillermo Campitelli, Fredrik Ullén, and Miriam A Mosing. 2016. Beyond born versus made: A new look at expertise. In *Psychology of learning and motivation*. Vol. 64. Elsevier, 1–55.

[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*. 3315–3323.

[21] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A Geometric Solution to Fair Representations. In *AIES*. 279–285.

[22] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *FAT**. 181–190.

[23] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *NIPS*. 325–333.

[24] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. 2020. InFoRM: Individual Fairness on Graph Mining. In *KDD*. 379–389.

[25] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm*. Oxford University Press.

[26] Derek Leben. 2017. A Rawlsian Algorithm for Autonomous Vehicles. *Ethics and Information Technology* 19, 2 (2017), 107–115.

[27] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum. Comput. Interact.* 3 (2019), 182:1–182:26.

[28] Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. 2009. Equality of Opportunity and Luck: Definitions and Testable Conditions, with an Application to Income in France. *Journal of Public Economics* 93, 11 (2009), 1189–1207.

[29] Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. 2009. Equality of Opportunity and Luck: Definitions and Testable Conditions, with an Application to Income in France. *Journal of Public Economics* 93, 11 (2009), 1189–1207.

[30] Michele Loi, Anders Herlitz, and Hoda Heidari. 2019. A Philosophical Theory of Fairness for Prediction-Based Decisions. *Available at SSRN 3450300* (2019).

[31] Alan Lundgard. 2020. Measuring Justice in Machine Learning. In *FAT**. 680.

[32] Koray Mancuhan and Chris Clifton. 2014. Combating discrimination using Bayesian networks. *Artificial Intelligence and Law* 22, 2 (2014), 211–238.

[33] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. 2020. Equality of Learning Opportunity in Personalized Recommendations. *CoRR* abs/2006.04282 (2020).

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *CoRR* abs/1908.09635 (2019).

[35] Richard E Nisbett. 2009. *Intelligence and how to get it: Why schools and cultures count*. WW Norton & Company.

[36] Richard E Nisbett, Joshua Aronson, Clancy Blair, William Dickens, James Flynn, Diane F Halpern, and Eric Turkheimer. 2012. Group differences in IQ are best understood as environmental in origin. *American psychologist* 67, 6 (2012), 503–504.

[37] Richard E Nisbett, Joshua Aronson, Clancy Blair, William Dickens, James Flynn, Diane F Halpern, and Eric Turkheimer. 2012. Intelligence: new findings and theoretical developments. *American psychologist* 67, 2 (2012), 130–159.

[38] Safiya Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce*. NYU Press.

[39] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[40] Dana Pessach and Erez Shmueli. 2020. Algorithmic Fairness. *CoRR* abs/2001.09784 (2020).

[41] Elizabeth Rapaport. 1981. Ethics and social policy. *Canadian Journal of Philosophy* 11, 2 (1981), 285–308.

[42] John Rawls. 1971. *A Theory of Justice*. Harvard University Press.

[43] John Rawls. 1999. *A Theory of Justice, Rev. Ed*. Harvard University Press.

[44] John Rawls. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.

[45] John E. Roemer. 2009. *Equality of Opportunity*. Harvard University Press.

[46] John E. Roemer and Alain Trannoy. 2015. Equality of Opportunity. In *Handbook of Income Distribution*. Vol. 2. Elsevier, 217–300.

[47] Ben Roshan. 2020. Campus Recruitment: Academic and Employability Factors Influencing Placement. https://www.kaggle.com/benroshan/factors-affecting-campus-placement/.

[48] Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall Press, USA.

[49] Tommie Shelby. 2003. Race and social justice: Rawlsian considerations. *Fordham L. Rev.* 72 (2003), 1697.

[50] Pavan Subhash. 2017. IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[51] Robert Taylor. 2004. Self-Realization and the Priority of Fair Equality of Opportunity. *Journal of Moral Philosophy* 1, 3 (2004), 333–347.

[52] Robert S. Taylor. 2009. Rawlsian Affirmative Action. *Ethics* 119, 3 (2009), 476–506. https://doi.org/10.1086/598170

[53] Pak-Hang Wong. 2020. Democratizing Algorithmic Fairness. *Philosophy & Technology* 33, 2 (2020), 225–244.

[54] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving Causal Fairness through Generative Adversarial Networks. In *IJCAI*. 1452–1458.

[55] Junzhe Zhang and Elias Bareinboim. 2018. Equality of Opportunity in Classification: A Causal Approach. In *NeurIPS*. 3671–3681.