Causal Modeling for Fairness in Dynamical Systems

Elliot Creager [1 2] **David Madras** [1 2] **Toniann Pitassi** [1 2] **Richard Zemel** [1 2]

Abstract

In many application areas—lending, education, and online recommenders, for example—fairness and equity concerns emerge when a machine learning system interacts with a dynamically changing environment to produce both immediate and long-term effects for individuals and demographic groups. We discuss causal directed acyclic graphs (DAGs) as a unifying framework for the recent literature on fairness in such dynamical systems. We show that this formulation affords several new directions of inquiry to the modeler, where causal assumptions can be expressed and manipulated. We emphasize the importance of computing interventional quantities in the dynamical fairness setting, and show how causal assumptions enable simulation (when environment dynamics are known) and off-policy estimation (when dynamics are unknown) of intervention on short- and long-term outcomes, at both the group and individual levels.

1. Introduction

How do we design fair policies for complex, evolving systems? Recently, the literature on fairness in dynamical systems has begun exploring the role of algorithmic systems in shaping their environments over time (Hashimoto et al., 2018; Lum & Isaac, 2016; Ensign et al., 2018). The key insight from these papers is that the repeated application of algorithmic tools in a changing environment can have fairness implications in the long-term distinct from those in the short-term.

However, the methods in this literature are quite disparate, with little overlap existing between various works in terms of modeling choices, goals, or assumptions. This lack of formal similarity is surprising, given that these papers are usually structurally alike: each proposes a dynamics model for a particular domain (e.g. lending (Mouzannar et al., 2019), hiring (Hu & Chen, 2018), recommendations (Bountouridis et al., 2019)), exposes unfairness that arises from long-term usage of some baseline policy, and then proposes a "fair" policy to mitigate some of these biases.

In this paper, we propose unifying the literature on fairness in dynamical systems via causal directed acyclic graphs (DAGs) (Pearl, 2009; Richardson & Robins, 2013). While causal DAGs have been used to study one-shot fair decision-making (Kusner et al., 2017; 2019; Kilbertus et al., 2017), they are uncommon in fairness settings involving *sequential* decisions. We show that several intuitive models of long-term unfairness are naturally expressed using causal DAGs. We also show that causal reasoning is useful for analyzing models and evaluating policies for these problems.

Our contributions are:

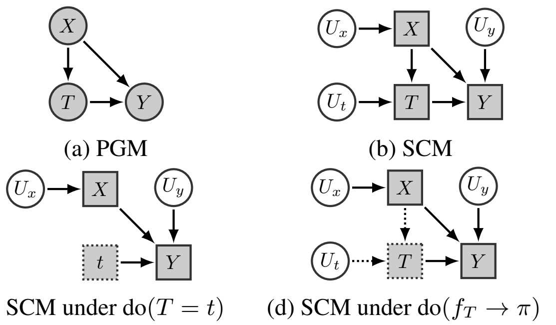
- We show that causal DAGs are a unifying framework for the literature on fairness in dynamical systems, reformulating examples from the literature using *structural causal models* and *policy interventions*.
- We demonstrate empirically that when environment dynamics are unknown, causal reasoning can help utilize observational data to improve off-policy estimation and learning.
- We show that if dynamics are known, causal DAGs serve as flexible simulators for analyzing policies and models, through extending and investigating model assumptions.

We proceed as follows. In Section 2, we introduce key background concepts of structural causal models and policy interventions. In Section 3, we demonstrate the application of causal DAGs to several key concepts in the fairness in dynamical systems literature. In Section 4 we discuss related work in fairness and causality. In Section 5 we empirically demonstrate that causal modelling can improve off-policy estimation and selection in a dynamical fairness problem, and in Section 6 we show how the explication of underlying causal assumptions enables model extension and analysis.[1]

[1]University of Toronto [2]Vector Institute. Correspondence to: Elliot Creager <creager@cs.toronto.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Code at `github.com/ecreager/causal-dyna-fair`



(a) PGM (b) SCM

(c) SCM under do$(T = t)$ (d) SCM under do$(f_T \to \pi)$

Figure 1: Treatment model expressed as PGM (1a), SCM (1b), and SCM under atomic (1c) and policy interventions (1d).

2. Background

2.1. Structural Causal Models

There are several ways of encoding causal assumptions in DAG form. In this paper, we focus on structural causal models (SCMs) (Pearl, 2009), which we overview here.[2] SCMs are similar to probabilistic graphical models (PGMs) (Koller & Friedman, 2009). They consist of nodes (random variables representing entities in the world) and edges (relationships between entities). However, whereas PGMs only specify a set of conditional independence relationships, SCMs specify a unique data generating process (analogously, a *particular* probability factorization, as opposed to the multiple isomorphic factorizations available in a PGM).

There are two types of nodes in SCMs. *Endogenous nodes* represent variables of interest within the model, while *exogenous nodes* are external random variables, representing the exclusive source of stochasticity induced on the observations (the endogenous nodes). The edges between nodes are deterministic functions called *structural equations*. Hence, a setting of the exogenous nodes corresponds to exactly one setting of the endogenous nodes. In Figure 1b, the dark squares are endogenous nodes, representing specific entities such as a credit score, a medical treatment, or a sensitive attribute. The light circles are exogenous nodes. Each endogenous node is the output of a structural equation, e.g. $T = f_T(X, U_T)$, $Y = f_Y(T, X, U_Y)$.

We can calculate causal quantities under a particular SCM by using the do-operator. Given the probability distribution implied by the SCM in Figure 1b (call the model $\mathcal{M}$ and the implied joint distribution p), we may wish to ask – "What would be the expected value of Y if T were set to 1?" The corresponding estimand can be denoted $\mathbb{E}_{p^{do(T=1)}}[Y]$. This differs from the more straightforward conditional probability $\mathbb{E}_p[Y|T = 1]$, which describes co-occurences of Y with $T = 1$ in the observed data. The expression $\mathbb{E}_{p^{do(T=1)}}[Y]$ indicates that expected value of Y is computed under a modified SCM which is specified by $do(T = 1)$; we denote this $\mathcal{M}^{do(T=1)}$, with the associated probability distribution $p^{do(T=1)}$. $\mathcal{M}^{do(T=1)}$ is intended to simulate a randomized experiment — if the true data-generating process is represented by $\mathcal{M}$, what would happen to the observed data if we forcibly change the data-generating process, so that $T = 1$ always? Graphically, $\mathcal{M}^{do(T=1)}$ is created by starting with $\mathcal{M}$ (Fig. 1b), removing from the graph all the incoming arrows to T (in this case, arrows originating from X and U_T), and setting $T = 1$ (yielding Fig. 1c). This is referred to as an *intervention*. Under certain conditions (Pearl, 2009), we can identify $\mathbb{E}_{p^{do(T=1)}}[Y]$ by using observational data generated by p to simulate sampling from $p^{do(T=1)}$. Intervening on the value of T in this way is an *atomic* intervention.

2.2. Policy Interventions and Off-Policy Evaluation

Alternatively, we can intervene directly on the structural equation governing T (Fig. 1d), resulting in model $\mathcal{M}^{do(f_T \to \pi)}$ with distribution $p^{do(f_T \to \pi)}$. When an intervention manipulates a structural equation corresponding to a decision maker's policy, we call this a *policy intervention*. Accordingly, we denote the structural equation under intervention as π to emphasize that it represents the decision maker's policy, distinct from the structural f_T present during the previous collection of observational data (which in turn could also be referred to as a policy, say $f_T = \pi_{\text{Hist}}$).

Consider an observational dataset generated by some historical policy π_{Hist}. We may wish to know the expected outcome for some policy $\pi \neq \pi_{\text{Hist}}$, but cannot directly test π in the world ourselves. This *off-policy evaluation* problem is particularly important in fairness contexts, where running a candidate policy in the world is frequently impossible due to ethical or practical reasons. In an SCM, off-policy evaluation constitutes estimating expected outcomes under a policy intervention. In the example from Fig. 1d, to estimate the expected value of Y under a new policy π, we specify our intervention with $do(f_T \to \pi)$, and the estimand would be $\mathbb{E}_{p^{do(f_T \to \pi)}}[Y]$. In general, to denote the expected value of a variable $\mathcal{U}$ under a target policy $\hat{\pi}$ which intervenes on a variable V, we write $\mathbb{E}_{p^{do(f_V \to \hat{\pi})}}[\mathcal{U}]$.

2.3. Benefits of Causal Graphs

While there are a variety of strategies for modeling in the causal inference literature (the potential outcomes framework of Rubin (2005) is a popular alternative[3]), we believe that causal *graphs* as pioneered by (Pearl, 2009) convey several benefits of particular interest in applications with

[2]Other overviews of various levels of detail can be found elsewhere (Pearl, 2009; Madras et al., 2019; Buesing et al., 2019) .

[3]We note that SWIGs (Richardson & Robins, 2013), while not the focus of this work, provide a graphical method to express potential outcomes that could also used to study long-term fairness.

fairness concerns. We outline these benefits below.

Visualization Many problems in long-term fairness have a large number of variables, and require collaboration across disciplines and with policy makers or regulators. Graphical structure allows for mathematical manipulation of many variables, and can convey basic assumptions to non-technical stakeholders.

Introspection Using causal language to be explicit about assumptions is useful for learning better policies (we discuss one such example in Sec. 5). Using a graph to convey the causal assumptions is a stylistic choice, but it facilitates the interrogation of complex assumptions (with many variables). Since the usefulness of causal inferences often rests assumptions that cannot be readily tested, it is especially important to hold these assumptions to a high degree of scrutiny; the use of graphs to convey causal assumptions could empower non-technical stakeholders to participate in this process.

Evaluation Causal graphs convey a number of methodological benefits, especially in improving off-policy evaluation (Sec. 5), enabling expressive simulation, and suggesting relevant sensitivity analyses (Sec. 6). Furthermore, encoding causal assumptions using graphical language exposes an underlying *computation graph*. Under mild assumptions, the topology of a computation graph can be used to programatically derive a large family of estimators for use in off-policy evaluation and gradient-based policy learning (Schulman et al., 2015; Weber et al., 2019). In the context of causal inference, graph topology can assist in determining the identifiability of policy interventions from observational data; see discussion of "dynamic treatment regimes" by Hernán & Robins (2020) for further detail.

3. Causal Interpretations of Dynamic Fairness Models

In this section, we demonstrate how SCMs present a unifying framework for the literature on fairness in dynamical systems. We focus on how causal mechanisms enable easier explication of underlying modeling assumptions, yielding insight into the component parts of the model, types of bias which could arise, and the effects of hypothetical interventions. Our aim is not to promote a particular dynamical model or fairness objective/constraint, either in general or for specific problem domains; rather we aim to provide a tool with which policymakers and practitioners alike can analyze a long-term unfairness problem. We discuss SCM formulations of three models of fairness in dynamical systems (see Appendix E for several more examples):

1. Fair-MDP: a motivating example showing how bias can arise in a generic sequential decision process.

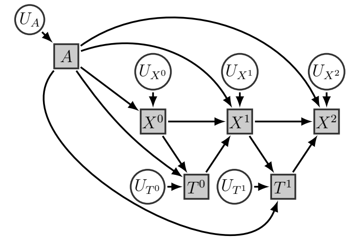


Figure 2: The Fair-MDP: a motivating example model for estimating questions of fairness in sequential decision-making problems.

2. Lending: Liu et al. (2018)'s single-step model of a loan application.
3. Repeated classification: Hashimoto et al. (2018)'s model of the dynamics of a changing population's preferences with unobserved sensitive attributes.

3.1. Fair-MDP: A Motivating Example

We begin by suggesting a minimal characterization of a sequential decision process in the fairness setting. Consider the following SCM (see Fig. 2), with the factorization:

$$\begin{aligned} A &= f_A(U_A) \\ X^0 &= f_{X^0}(A, U_{X^0}) \\ T^k &= f_T(X^k, A, U_{T^k}), k = 0 \dots K \\ X^{k+1} &= f_X(X^k, T^k, A, U_{X^{k+1}}), k = 0 \dots K \end{aligned} \tag{1}$$

This is similar to a Markov Decision Process (MDP), in that the key elements are states (X), actions (T), a policy (f_T) and a transition function (f_X). However, we note that it is not fully Markovian — the sensitive attribute persists across states, affecting all aspects of the problem. This aligns with standard fairness intuitions, since the sensitive attribute is generally considered to be somewhat holistic and immutable by T. We denote this model the *Fair-MDP*, since it becomes an MDP when we condition on the sensitive attribute, and the inclusion of this attribute permits fairness considerations.[4] We can think of X as some feature of an individual, which our policy is aiming to maximize, and consider the final X^k in the sequence as the reward.

We can use this model to examine different fairness issues in the sequential setting. For instance, consider the issue of feeback loops (Ensign et al., 2018;

[4] In general, MDPs are typically used to define all relevant variables as part of a "state", and methods from this literature can be applied to long term fairness problems. Causal graphs enable fine-grained modeling of the dynamics with a state (see Fig. 12 in Appendix E for one such example), which can be practically useful in long-term fairness problems, e.g. in improving off-policy evaluation (Sec. 5).

Lum & Isaac, 2016). Suppose that the initial feature distribution $P(X^0|A)$ is uneven: $\mathbb{E}[X^0|A = 1] > \mathbb{E}[X^0|A = 0]$. Additionally, suppose a threshold policy is applied, with $T^k = f_T(X^k, \cdot, \cdot) = \mathbb{1}(X^k > \tau)$ and that the application of the treatment causes X to increase: $\mathbb{E}(X^{k+1}|X^k, T^k = 1) > X^k$ (and $T^k = 0$ causes the opposite effect). Then, we might expect to see a *feedback loop*, as observed in Ensign et al. (2018), where one group's average reward increases continuously over time, and the other group's decreases.

Off-policy estimation for a policy π in this model amounts to estimating $\mathbb{E}_{p^{do(f_T \to \pi)}}[X^K]$. We note that this is a nontrivial problem — if we only observe data generated by some historical policy $\pi_H \neq \pi$, then the values of X^K under the actions that π would have taken may not be available in our data. In this case, the naive estimator $\mathbb{E}_p[X^K]$ will be biased. We return to the off-policy estimation question in Section 5, with a causal approach.

3.2. Lending

We turn to the model from Liu et al. (2018), which examines threshold-based classification in general, but with specific focus on the lending setting. Our SCM formulation of this model can be seen in Figure 4a. In this model, a person with group membership (a.k.a. sensitive attribute) A receives a credit score X, and applies to a bank for a loan. The bank makes a binary decision T about whether to award the loan using the policy f_T. The binary potential outcome Y is realized, which is converted to institutional profit or loss only if $T = 1$.[5] Finally, the applicant's credit score is modified to $\tilde{X}$ (increased on repayment, decreased on default, static if $T = 0$).[6] The bank's utility is measured through their profit $\mathcal{U}$ (a sum over the individual profits u) as well as the expected score change Δ_j, representing the average change in credit score after one time-step among members of group $A = j$. Varying the loan policy can achieve different values of $\mathcal{U}, \Delta_j$, resulting in outcomes with different fairness properties.

Liu et al. (2018) consider the effect of various threshold policies for loan assignment under this model, namely the expected values of $\mathcal{U}$ and Δ for some policies with group-specific thresholds $\tau \triangleq (\tau_0, \tau_1)$ that offer loans to applicants of group j with score X if and only if their credit score $X > \tau_j$. They show that different thresholds satisfy different criteria: maximum profit (MaxProf), demographic parity (DemPar), and equal opportunity (EqOpp). In the language of our paper, comparing threshold policies is done through policy evaluation and intervention. Denoting by π_τ a threshold policy per group τ, these results can be phrased with the tool of policy intervention: we evaluate the policy π_τ by estimating the quantities $\mathbb{E}_{p^{do(f_T \to \pi_\tau)}}[\mathcal{U}]$ and $\mathbb{E}_{p^{do(f_T \to \pi_\tau)}}[\Delta_j]\ \forall j$, for various τ computed under different fairness criteria. We discuss off-policy evaluation in this model in Section 5.

This SCM interpretation suggests several potential extensions, such as evaluating outcomes over multiple steps or adding extra actors to the model. We discuss these in detail in Section 6, where we provide a case study of this SCM.

3.3. Repeated Classification

Finally, we examine the repeated classification setting discussed by Hashimoto et al. (2018), presented in SCM form in Figure 3. The model is fairly general, and the authors discuss several domains where it could apply (e.g. speech recognition, text auto-completion). A binary classifier with parameters θ is repeatedly trained on a population of individuals with features X and labels Y. The population distribution is a mixture of components $P = \sum_k \alpha_k P_k$, where each of the k demographic groups has proportion α_k (with $\sum_k \alpha_k = 1$) and a unique distribution over the input-output pairs $P_k(X, Y)$. Group memberships (i.e. cluster assignments) $Z \in [1 \dots k]$ are not observed.

The key idea is that the group distributions P_k remain static over time, but their relative proportions α_k change dynamically in response to the classifier performance on the k-th group. At the t-th step, the classifier is trained on the overall population $\{(X_i^t, Y_i^t)\}$, yielding classifier parameter θ^t

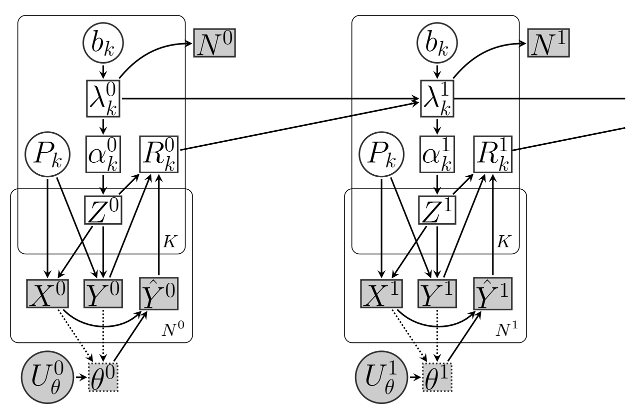


Figure 3: SCM for the repeated loss minimization model discussed by Hashimoto et al. (2018). The dotted arrows highlight the policy as the learning algorithm that produces parameters θ^t, which in turn affect the predictions $\hat{Y}_i^t$. See Table 2 for explanation of all symbols and text for description.

[5] Therefore this model does not capture a notion of opportunity loss for *not* extending a loan to applicants who *are* qualified.

[6] Likewise, the applicant's score does not change in the absence of a loan; this assumption may be inaccurate, since *not* receiving a loan could create additional financial issues for the applicant.

and predictions[7] $\hat{Y}^t$. At each step, some subjects choose to stay in the population, some choose to leave, and some new subjects are added to the pool. In particular, the Poisson parameter λ_k (proportional to mixing coefficient α_k) is computed as a function of the per-group risk R_k. Misclassified subjects are more likely to leave, so under-served groups shrink over time. The authors coin this phenomenon as *disparity amplification*. Interestingly, disparity amplification can *improve* the overall loss/accuracy since the shrinking minority group (whose accuracy may be decreasing) contributes less to these global metrics as time proceeds. To mitigate disparity amplification, Hashimoto et al. (2018) propose a robust optimization technique that seeks low loss for worst-case group assignments Z (assuming a minimum group size).

This SCM suggests several interesting interventions:

1. Intervention on *latent dynamics*: $\text{do}(f_\lambda \rightarrow \hat{f}_\lambda)$ represents an intervention on population dynamics, which we could use to test how policies affect the entry and exit of various groups from the environment over time. $\text{do}(b_k = \hat{b}_k)$ is a simple atomic intervention of a similar flavor, which changes the expected number of individuals entering each group at a given time step.

2. Intervention on *group distributions*: $\text{do}(P_k \rightarrow \hat{P}_k)$ shifts the distribution over input-output pairs for group k, which could be carried out at one or every time step.

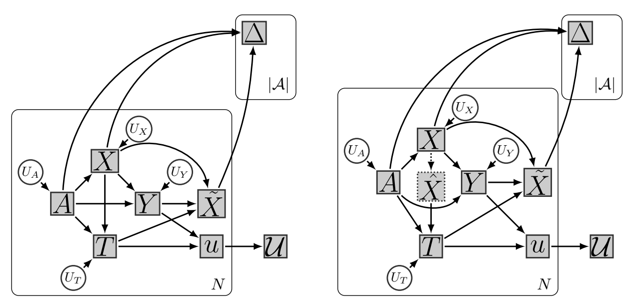
We do not present experiments on this model, but include it to suggest the types of analyses and extensions possible for SCMs with increased complexity. See Appendix E for more sophisticated models from the fairness in dynamical systems literature represented as SCMs.

4. Related Work

Dynamical Fairness There has been work on modeling the long-term dynamics of fairness in a range of potential domains. Recently, the first paper to bring these issues to light was Lum & Isaac (2016), discussing the bias feedback loops which could arise in predictive policing systems, with follow-up work by Ensign et al. (2018). Domains such as hiring (Hu & Chen, 2018), loans (Mouzannar et al., 2019), and recommender systems (Hashimoto et al., 2018; Bountouridis et al., 2019) have also been explored in this way. Other related explorations have dealt with short-term dynamics (Liu et al., 2018) and strategic actions (Hu et al., 2019; Milli et al., 2019). There is also a line of work studying the long-term effects of affirmative action, with some classic works from the economics literature (Coate & Loury, 1993; Foster & Vohra, 1992), and more recent computer science focused work (Kannan et al., 2019). On the theoretical side, several general algorithms for improved fairness in sequential decision-making have been characterized, with work discussing bandits (Joseph et al., 2016), reinforcement learning (Jabbari et al., 2017), and importance sampling estimators (Doroudi et al., 2017). The work of D'Amour et al. (2020)—which most closely relates to ours—studies long-term outcomes for existing fair ML methods, emphasizing *agents* and *environments* as modeling primitives. Our contributions can be seen as complementary, emphasizing the role of *causal* modeling primitives within a dynamical system, both in terms of estimation from observational trajectories, and building expressive simulators for evaluating agents and environments.

Causality Causal modeling has been used in a variety of non-dynamic fair machine learning approaches. Work on counterfactual fairness (Kusner et al., 2017) has considered fairness definitions which encourage models to treat examples similarly to hypothetical situations where they were from the other group. Some other works focus on learning fair policies from biased observational data (Madras et al., 2019; Kusner et al., 2019) or on learning decision rules which follow only causal paths deemed to be non-discriminatory (Kilbertus et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019). Another line of work interprets previously proposed fairness criteria from a causal perspective (Zhang & Bareinboim, 2018a;b).

Outside of fairness, Everitt et al. (2019) propose using influence diagrams as a framework for understanding safety in AI systems.



(a) . Our SCM formulation of the one-step dynamics.

(b) An extension emphasizing the role of the credit bureau.

Figure 4: Causal interpretation of lending scenario from Liu et al. (2018). See Sections 3.2, 5 and 6 for discussion, and Table 1 in Appendix D for symbol legend.

5. Off-Policy Evaluation and Selection

Given historical observations, how can we estimate the real-world impact of deploying a *new* policy (e.g. one that incorporates fairness constraints)? This question motivates one of

[7] Using held-out data for the predictions is expressible via a small change to the SCM.

the key tasks required for improving long-term ML fairness: *off-policy evaluation*. As noted in Sec. 2, here we must rely on *observational data* since it is often unethical or unsafe to test candidate policies in the world (e.g. an A/B test). In this section, we demonstrate empirically that causal reasoning improves off-policy evaluation from observational data.

In this experiment, we consider a scenario where the bank has historical data from a profit-maximizing policy (MAXPROF) and wishes to learn and estimate the quality of an equal opportunity policy (EQOPP) before deploying it (the off-policy estimation/learning problem). We use the lending setting of Liu et al. (2018) under our SCM interpretation[8] (see Figure 4a for depiction and Appendix C for full specification). The key (non-trivial) structural equations of the SCM are:

$$\begin{aligned} X &= f_X(A, U_X) \\ T &= f_T(X, A, U_T) \\ Y &= f_Y(X, A, U_Y) \end{aligned} \tag{2}$$

which are the feature distribution, the historical treatment policy, and the outcome distribution, respectively. The change in individual score c, the bank's utility u, and the next-step score $\tilde{X}$, are simple functions of the other variables: $(c, u) = (c_+, u_+)$ if $Y = 1$ or (c_-, u_-) if $Y = 0$, and $\tilde{X} = X + c$ (for constants $c_+, u_+ > 0; c_-, u_- < 0$). As in Liu et al. (2018), we focus on threshold policies, which are defined by group-specific thresholds $\tau \triangleq (\tau_0, \tau_1)$ that offer loans to applicants of group j with score X if and only if their credit score $X > \tau_j$.

5.1. Procedure

In order to compute good thresholds τ_j for various lending policies (maximum profit, equal opportunity, etc.), Liu et al. (2018) make a very strong assumption in their method: that these underlying dynamics parameters $(f_X, f_T, f_Y, c_+, c_-, u_+, u_-)$ of the system are known. This is stronger than just assuming the causal structure, as we do in Fig. 4a. The causal structure implies the general functional form for the data generating process. However, Liu et al. (2018) assume not just the form but that the function parameter values are known. In practice, these functions will rarely be known, and must be estimated from observational data. Therefore any off-policy selection or learning hinges on the quality of these estimates.

Some of these unknown parameters (e.g. u_+, u_-, f_T) are easy to estimate from data. However, one in particular is difficult: the outcome function $Y = f_Y(X, A, U_Y)$. To understand why estimating f_Y from data is difficult, we must note that Y is a *causal* quantity. Specifically, Y is a *potential outcome* (Rubin, 2005): it is the probability of a person repaying a loan *were they to receive one*.[9] Estimating Y is difficult because it is often missing: we only observe Y when a loan was given in the observational data. Therefore, straightforward estimates may be biased or high variance.

This difficulty of estimating Y propagates into the rest of the problem; u and Δ have the same issues: they are potential outcomes, only observed when the treatment is given ($T = 1$). Therefore, choosing the policy thresholds—which involves estimating (u, Δ)—is inherently a causal problem.

Given a policy π, we focus on computing an off-policy estimator $\mathcal{E}(\pi) \approx \mathbb{E}_{p^{do(f_T \to \pi)}}[u]$. A simple estimator can be derived via regression: first learn a function to approximate $f_{\text{Reg}}(X, A) \approx \mathbb{E}_{p^{\text{obs}}}[u|X, A]$ in the observational data; then apply this regression for every individual where π suggests giving the treatment: $\mathcal{E}_{Reg}(\pi) = \mathbb{E}_{p^{\text{obs}}(X,A)}[f_{\text{Reg}}(X, A)|\pi(X, A) = 1]$. This is a natural baseline in the absence of causal reasoning.

However, we can further improve this estimator. As noted previously, u is missing from the observational data in a biased way. Therefore, we can approach the off-policy estimation problem as a missing data problem — an area for which causal inference has developed a number of tools. Crucially, the set $\{X, A\}$ satisfies the *backdoor criterion* from T to u in the SCM (see Fig. 4a). This justifies[10] the use of a doubly robust estimator as presented by Zhang et al. (2012), an estimator that combines a regression-based and an inverse-propensity estimator (Bang & Robins, 2005) to reduce bias and variance.[11] With $C_i = \mathbb{1}[\pi(X_i, A_i) = T]$, the estimator is

$$\begin{aligned} \mathcal{E}_{DR} = \frac{1}{n} \sum_{i=1}^{n} \Big[&\frac{C_i(\pi) u_i}{P(C_i(\pi) = 1|X_i, A_i)} \\ &- \frac{C_i(\pi) - P(C_i(\pi) = 1|X_i, A_i)}{P(C_i(\pi) = 1|X_i, A_i)} f_{Reg}(X_i, A_i) \Big]. \end{aligned}$$

We can use an analogous estimator for Δ, where the same

[8] Our aim in this Section is to demonstrate the additional capabilities of causal modeling in this context, and not to adjudicate on whether this choice of dynamics model is correct or appropriate to the lending setting. In decision making problems with material consequences for individuals and groups (such as lending), care should be taken when incorporating sensitive group information into a causal model. For example, the generative process studied here includes a causal link from sensitive attribute (which represents race) to credit score. Although this is consistent with the original paper and captures the statistics of the dataset, it could also be subject to criticism around modeling race as a cause of social position rather than being socially ascribed; see Benthall & Haynes (2019) and Hanna et al. (2020) for further discussion.

[9] Using the notation of Rubin (2005), we could denote it as Y_1.

[10] We also rely on the assumptions of *overlap* (in this case requiring a mildly stochastic historical policy) and *consistency*. While these are typical in the causal inference literature, they may be difficult to verify from observational data in some settings. See Appendices B and C for further discussion.

[11] The doubly robust estimator can also be interpreted as applying the regression estimator as a control variate to the importance sampling estimator; see Dudík et al. (2011) for discussion.

backdoor criterion holds. To summarize, we took the following steps to derive the causal estimator:

1. We first recognized that u was implicitly a causal quantity (a potential outcome).
2. We next examined the SCM to identify a set of variables which satisfied the backdoor criterion between T and u.
3. We finally applied a specialized estimator from the causal inference literature to improve performance.

5.2. Results

We generate observational data from the SCM in Figure 4a, under a MaxProf threshold policy. We then consider a new policy π_τ with per-group thresholds $\{\tau_j\}$ as its two parameters. We compute the estimators $\mathcal{E}_{Reg}(\pi_\tau)$ and $\mathcal{E}_{DR}(\pi_\tau)$ for varying values of these thresholds. Figure 5 shows that the causally motivated estimator $\mathcal{E}_{DR}$ achieves lower off-policy estimation error on both sensitive groups, across the threshold range. Note the high estimation error of the baseline $\mathcal{E}_{Reg}$ for low values of τ. This is because the historical policy typically does not award loans to applicants with low scores, meaning there are fewer data available for the regression.

Ultimately, the goal of estimating these quantities is to improve policy learning. We can formulate an objective which trades off between utility and an equal opportunity term $\delta_{EqOpp} = |P(T=1|Y=1, A=0) - P(T=1|Y=1, A=1)|$. The overall objective is $\mathcal{V}_\pi = \mathcal{U} - \lambda\delta_{EqOpp}$. We hope to maximize this, with some hyperparameter $\lambda \in \mathbb{R}$ governing the tradeoff. We note that estimating δ_{EqOpp} itself presents a challenging causal problem, since Y is frequently unobserved. See Appendix B for details on this estimation problem and the rest of this experiment.

Using the estimators presented above, we can construct an off-policy estimate of $\mathcal{V}_\pi$. We search over the space of two-threshold policies (one threshold per group) to find the policy with the highest off-policy estimate of the objective on a validation set. We then calculate the true value of $\mathcal{V}_\pi$ on

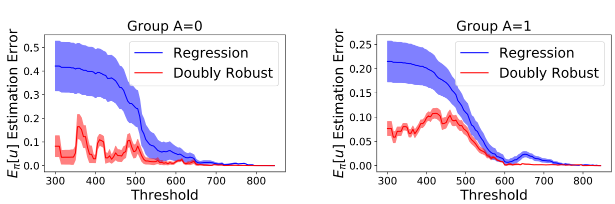


Figure 5: Comparing error of $\mathcal{E}_{Reg}$ and $\mathcal{E}_{DR}$ (regression and doubly robust) for off-policy estimation of $\mathbb{E}_\pi[u]$ from observational data. The x-axis represents single threshold policies. Estimation for Δ yields similar results, since both variables are linear in Y.

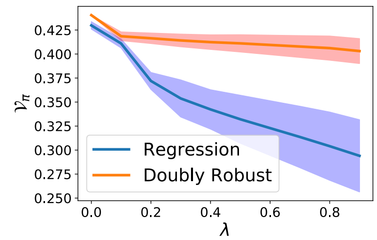


Figure 6: Test set value of a fairness-utility objective using the two off-policy estimators. Hyperparameter λ governs the tradeoff. Higher values of the objective $\mathcal{V}_\pi$ are better.

a held-out test set, using the SCM simulator (as visualized in Figure 4a and specified fully in Appendix C) to generate the true potential outcomes. The estimator $\mathcal{E}_{DR}$ that more fully incorporates causal reasoning in the parameter estimation finds a better objective value, ultimately yielding an improved policy (see Fig. 6). We emphasize that this improvement requires assumptions about causal structures, but not precise knowledge of the system dynamics.

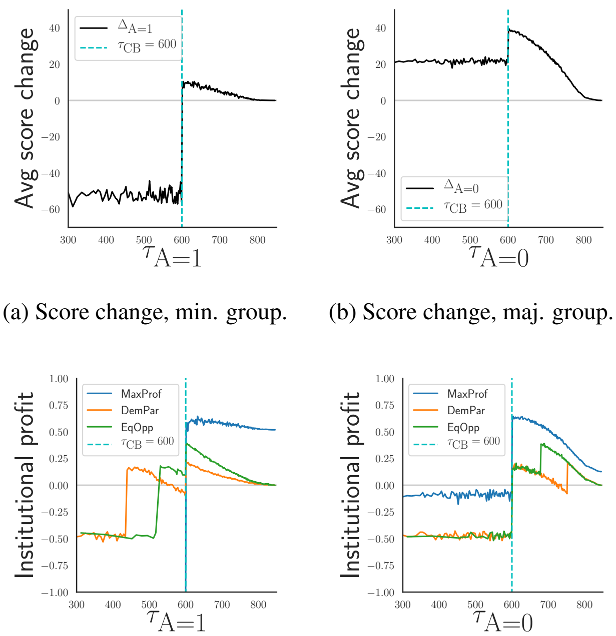
6. Extensions in Lending via Intervention

We now investigate the setting where both causal structure *and* dynamics are known (returning to the assumptions made by Liu et al. (2018)), and emphasize the role of *interventions* in building expressive simulators for dynamical fairness settings. Thus we carry out "on policy" evaluations that sample from the SCM directly. SCMs enable clearer explication of underlying causal assumptions. This means the framework is flexible: novel policy interventions extend our model by modifying existing assumptions, or testing our reliance on the assumptions we have already made. We give two such examples, measuring: (a) the interaction of the lender with other agencies; and (b) the sensitivity of long-term outcomes to the lender's modeling assumptions.

6.1. Multi-actor Experiments

Intervention by credit bureau Liu et al. (2018) conduct experiments based on statistics of FICO credit scores assigned by the credit bureau TransUnion (Reserve, 2007). We note that these credit score decisions themselves constitute a policy; and moreover, the language of interventions in the SCM framework allows us to characterize decisions made by the *credit bureau* (rather than the bank) using the same fairness and profit metrics as before.[12] The credit bureau enters the SCM by reinterpreting X_i as *features* related to creditworthiness of an individual, then introducing

[12]Note that recent changes by the credit scoring bureau Fair Isaac Corp. (https://www.wsj.com/articles/fico-changes-could-lower-your-credit-score-11579780800) can be characterized as such an intervention.



(a) Score change, min. group. (b) Score change, maj. group.

(c) Profit as fn. of min. thresh. (d) Profit as fn. of maj. thresh.

Figure 7: Policy evaluation under credit bureau intervention $\hat{f}_{\hat{X}}(X) = \min(X, \tau_{CB})$ with $\tau_{CB} = 600$. Group score change—formally $\mathbb{E}_{p^{do(f_{\hat{X}} \to \hat{f}_{\hat{X}}, f_T \to \hat{f}_T)}}[\Delta_j]$—and institutional profits—$\mathbb{E}_{p^{do(f_{\hat{X}} \to \hat{f}_{\hat{X}}, f_T \to \hat{f}_T)}}[\mathcal{U}]$—are shown as functions of the two group thresholds $\{\tau_j\}$ under several fairness constraints.

$\hat{X}_i = f_{\hat{X}}(X_i)$ as a *score* that is deterministically computed by the agency from the features (See Fig. 4b). When $f_{\hat{X}}$ is the identity function, we recover the original model. Policy evaluation under double intervention $\mathcal{M}^{\mathrm{do}(f_T \to \hat{f}_T, f_{\hat{X}} \to \hat{f}_{\hat{X}})}$ captures the sensitivity of the bank's decisions to the decisions of the credit bureau (and vice versa).

Results Figure 7 shows the effect on the average utility $\mathbb{E}[\mathcal{U}]$ and average per-group score change $\mathbb{E}[\Delta_j]$ of a simple policy intervention by the credit bureau. The intervention involves the bureau setting the minimum score to 600 for all applicants via the structural equation $\hat{f}_{\hat{X}}(X) = \min(X, 600)$. This intervention is unlikely in the real world because it contradicts the profit incentives of the bureau, which encourage well-calibrated scores. Nevertheless, it coarsely captures a potential scenario where an actor besides the bank seeks to encourage fair outcomes in a group-blind way, since under the new scoring policy minority applicants are more likely to receive loans. However, we see in Figure 7a that the average group outcome for protected applicants ($A = 1$) worsens when the bank's group threshold $\tau_{A=1}$ is below 600, since in this case its policy offers loans to individuals who have good scores on paper but are unlikely to repay the loans. Interestingly, the expected profit (Figure 7b) under credit bureau intervention differs depending on the fairness criteria of the bank. This is because each fairness criteria differently constrains the relationship between the two thresholds $\{\tau_{A=0}, \tau_{A=1}\}$ (the protected group is $A = 1$), so the choice of fairness criteria implicitly sets how many applicants with boosted scores ($X < 600$, thus $\hat{X} = 600$) are selected for loans. DEMPAR is more sensitive to the credit bureau intervention than EQOPP; it obeys a stricter fairness constraint and offers more loans to applicants with boosted scores (who are are unlikely to repay, and disproportionately belong to the minority group).

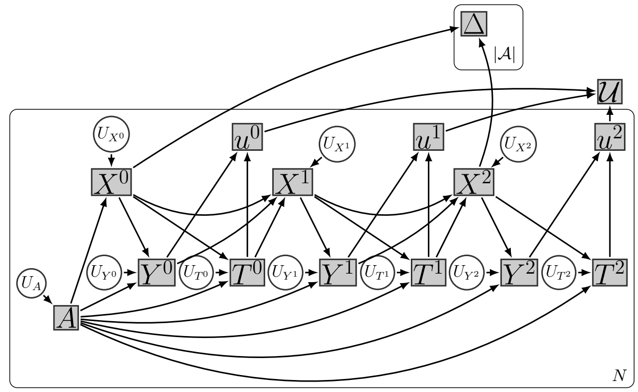


Figure 8: Phrasing the model from Liu et al. (2018) as an SCM enables a multi-step extension for measuring long-term impacts, e.g., in the two-step version shown here.

6.2. Sensitivity Analysis of Long-term Outcomes

Sensitivity analysis (Rosenbaum, 2014; Saltelli et al., 2008)—the task of measuring how sensitive a system's output is to its various assumptions—is critical when engaging in a complex modeling task. Since causal language makes structural modeling assumptions explicit, it is a natural match for sensitivity analysis. Questions of robustness are particularly important in long-term, dynamic modeling, since small assumptions errors can have large effects downstream when propagated over time. In this section, we show how to conduct a long-term sensitivity analysis of the Liu et al. (2018) model with SCMs, probing how sensitive proposed policies may be to underlying causal assumptions. We cast the sensitivity analysis as an on-policy evaluation under an intervention that accounts for model mismatch.[13]

Long-term impacts Given a policy whose one-step effect is purportedly fair, what can we say about its longer-term impacts? The modularity of the SCM formulation allows us to easily estimate these effects. For example, the struc-

[13]"Mismatch" refers here to structural equations with misspecified functional forms, not incorrect assumptions of causal structure.

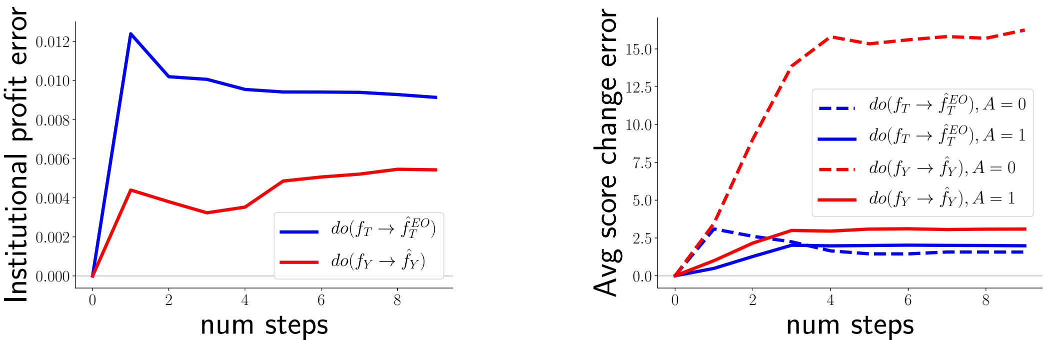


Figure 9: Evaluating multi-step policy robustness to distribution shift for various choice of intervention distribution q. Sensitivity of institutional utility—formally $|\mathbb{E}_q[\mathcal{U}] - \mathbb{E}[\mathcal{U}]|$—and sensitivity of group avg. score change—formally $|\mathbb{E}_q[\Delta_j] - \mathbb{E}[\Delta_j]|$—are shown as a function of steps. Expected profit is relatively robust to both interventions, whereas the expected per-group score changes are relatively more sensitive to these interventions.

tural equation $\tilde{X} = f_{\tilde{X}}(X, Y, T)$ can be modified to the recursive update $X^{t+1} = f_{X^{t>0}}(X^t, Y^t, T^t)$. Note that X^t (which does not directly depend on A) is only computed in this way for steps $t > 0$, since the original scores X^0 are sampled from $p(X^0|A)$. On the other hand, since T^t and Y^t still depend on $A \; \forall \; t$, we see that group membership does indeed have a long-term influence on the outcomes and score trajectories for individuals.

Results We conduct our sensitivity analyses as simulations of policy interventions under varying underlying model assumptions. We analyze the sensitivity of the EQOPP policy to two forms of model mismatch. In the first, do$(f_T \rightarrow \hat{f}_T^{EO})$ recomputes the per-group thresholds under the EQOPP constraint, but using *incorrect* statistics from the credit bureau. In particular, the marginal $p(Y|X)$ was used for both group's repayment probabilities rather than the correct $p(Y|X, A)$. The second intervention do$(f_Y \rightarrow \hat{f}_Y)$ is more severe, as $p(Y|X)$ is used to *sample* potential outcomes Y rather than just set the thresholds within f_T. We measure error under each intervention relative to the "ground truth" baseline where the correct potential outcome distributions are used to set thresholds and sample data. We measure how these errors compound over time (Figure 9). We observe the institutional profits are surprisingly robust to both forms of intervention, while the per-group outcomes are more sensitive to these interventions, especially to do$(f_Y \rightarrow \hat{f}_Y)$. These results indicate that our policies are particularly sensitive to assumptions around outcome prediction for the sensitive groups.

7. Discussion

In this paper, we discuss causal modelling as a unifying framework for the literature on fairness in dynamical systems. We demonstrate that in the realistic situation where the dynamics parameters of these systems are not known, causal models are helpful for estimating these parameters, and evaluating and learning policies in an off-policy manner from historical data. Additionally, we show how a causal model can be used as a simulator when the parameters *are* known, and how the modularity of the SCM framework is helpful for both expressing natural extensions existing work from the literature, and running long-term sensitivity analyses of policy decisions.

Since a causal DAG can be thought of as an expressive simulator, standard tools for optimization/learning in computation graphs (Schulman et al., 2015) can be brought to bear in order to *learn* policies that capture optimal rewards *across many interventional settings*. Using gradient estimators to learn policies in this setting holds promise in scaling to high dimensional datasets, which we leave for future work.

Acknowledgements

We thank Kuan-Chieh Wang, Eleni Triantafillou, Will Grathwohl, James Lucas, Robert Adragna, and John Miller for providing feedback, and the reviewers for their helpful suggestions.

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners.

References

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Benthall, S. and Haynes, B. D. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 289–298, 2019.

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., and Hauff, C. Siren: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 150–159. ACM, 2019.

Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.-B., and Heess, N. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Representation Learning*, 2019.

Coate, S. and Loury, G. C. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pp. 1220–1240, 1993.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Doroudi, S., Thomas, P. S., and Brunskill, E. Importance sampling for fair policy selection. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pp. 160–171, 2018.

Everitt, T., Ortega, P. A., Barnes, E., and Legg, S. Understanding agent incentives using causal influence diagrams, part i: single action settings. *arXiv preprint arXiv:1902.09980*, 2019.

Foster, D. P. and Vohra, R. V. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512, 2020.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Hernán, M. A. and Robins, J. M. Causal inference: what if. *Boca Raton: Chapman & Hill/CRC*, 2020, 2020.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268. ACM, 2019.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1617–1626. JMLR. org, 2017.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

Kannan, S., Roth, A., and Ziani, J. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 240–248. ACM, 2019.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kusner, M., Russell, C., Loftus, J., and Silva, R. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*, pp. 3591–3600, 2019.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Liu, L., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3156–3164, 2018.

Lum, K. and Isaac, W. To predict and serve? *Significance*, 13(5):14–19, 2016.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 349–358. ACM, 2019.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239. ACM, 2019.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368. ACM, 2019.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Nabi, R., Malinsky, D., and Shpitser, I. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Reserve, U. F. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Washington, DC: Board of Governors of the Federal Reserve System*, 2007.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Rosenbaum, P. R. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

Schulman, J., Heess, N., Weber, T., and Abbeel, P. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pp. 3528–3536, 2015.

Weber, T., Heess, N., Buesing, L., and Silver, D. Credit assignment techniques in stochastic computation graphs. *arXiv preprint arXiv:1901.01761*, 2019.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Zhang, J. and Bareinboim, E. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pp. 3671–3681, 2018a.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b.