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 (Bontouridis 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.

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


1Code at github.com/ecreager/causal-dyna-fair
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. 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 $M$ 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 $E_{p^{do(T=1)}}[Y]$. This differs from the more straightforward conditional probability $E_p[Y | T = 1]$, which describes co-occurrences of $Y$ with $T = 1$ in the observed data. The expression $E_{p^{do(T=1)}}[Y]$ 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 $E_{p^{do(T=1)}}[Y]$ 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 $\mathcal{M}^{do(f_T \rightarrow \pi)}$ with distribution $p^{do(f_T \rightarrow \pi)}$. 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 $\pi$ 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 $\pi_{\text{Hist}}$. We may wish to know the expected outcome for some policy $\pi \neq \pi_{\text{Hist}}$, but cannot directly test $\pi$ 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 $\pi$, we specify our intervention with do$(f_T \rightarrow \pi)$, and the estimand would be $E_{p^{do(f_T \rightarrow \pi)}}[Y]$. In general, to denote the expected value of a variable $\mathcal{U}$ under a target policy $\pi$ which intervenes on a variable $V$, we write $E_{p^{do(V \rightarrow \pi)}}[\mathcal{U}]$.

### 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), we believe that causal graphs as pioneered by (Pearl, 2009) convey several benefits of particular interest in applications with

---

2We 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.

#### 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:

\[
A = f_A(U_A) \\
X^0 = f_{X^0}(A, U_{X^0}) \\
T^k = f_T(X^k, A, U_{T^k}), k = 0 \ldots K \\
X^{k+1} = f_X(X^k, T^k, A, U_{X^{k+1}}), k = 0 \ldots K 
\]

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 feedback loops (Ensign et al., 2018;...
We turn to the model from Liu et al. (2018), which examines loan with different fairness properties. A loan could create additional financial issues for the applicant. Not extending a loan to applicants who are qualified.

Lum & Isaac, 2016). Suppose that the initial feature distribution \( P(X^0|A) \) is uneven: \( E[X^0|A = 1] > E[X^0|A = 0] \). Additionally, suppose a threshold policy is applied, with \( T^k = f_T(X^k, \cdot, \cdot) = 1(X^k > \tau) \) and that the application of the treatment causes \( X \) to increase: \( E(X^{k+1}|X^k, T^k = 1) > X^k \) (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 \( \pi \) in this model amounts to estimating \( E_{p^\text{old}}[\pi^\text{new}][X^K] \). We note that this is a non-trivial 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 \( \pi \) would have taken may not be available in our data. In this case, the naive estimator \( E_p[X^K] \) 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 outcome \( Y \) is realized, which is converted to institutional profit or loss only if \( T = 1 \). Finally, the applicant’s credit score is modified to \( X' \) (increased on repayment, decreased on default, static if \( T = 0 \)). The bank’s utility is measured through their profit \( U \) (a sum over the individual profits \( u \)) as well as the expected score change \( \Delta_j \), 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 \( U, \Delta_j \), 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 \( U \) and \( \Delta \) for some policies with group-specific thresholds \( \tau = (\tau_0, \tau_1) \) 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 \( \pi_\tau \) a threshold policy per group \( \tau \), these results can be phrased with the tool of policy intervention: we evaluate the policy \( \pi_\tau \) by estimating the quantities \( E_{p^\text{old}}[\pi_\tau][U] \) and \( E_{p^\text{old}}[\pi_\tau][\Delta] \), for various \( \tau \) 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 \( \theta \) 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 \( \alpha_k \) (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) \( Z \in [1 \ldots k] \) are not observed.

The key idea is that the group distributions \( P_k \) remain static over time, but their relative proportions \( \alpha_k \) 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 \( \{(X^t_i, Y^t_i)\} \), yielding classifier parameter \( \theta^t \)

![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 \( \theta^t \), which in turn affect the predictions \( \hat{Y}^t_i \). See Table 2 for explanation of all symbols and text for description.](image-url)
and predictions $\hat{Y}^t$. 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 $\lambda_k$ (proportional to mixing coefficient $\alpha_k$) 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. $\text{do}(b_k \rightarrow \hat{b}_k)$ 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.

---

![Diagram](https://via.placeholder.com/150)

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

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

---

**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
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 (EOPP) before deploying it (the off-policy estimation/learning problem). We use the lending setting of Liu et al. (2018) under our SCM interpretation (see Figure 4a for depiction and Appendix C for full specification). The key (non-trivial) structural equations of the SCM are:

\[
X = f_X(A, U_X) \quad (2)
\]

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 \((c, u )= (c_-, u_-)\) 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 \( \tau = (\tau_0, \tau_1) \) 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 \( \tau_j \) for various lending policies (maximum profit, equal opportunity, etc.), Liu et al. (2018) make a very strong assumption in their method: that the 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. 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 have high variance.

This difficulty of estimating \( Y \) propagates into the rest of the problem; \( u \) and \( \Delta \) 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 \((u, \Delta)\)—is inherently a causal problem.

Given a policy \( \pi \), we focus on computing an off-policy estimator \( \mathcal{E}(\pi) \approx E_{\pi}(u|f_T = \tau) \). A simple estimator can be derived via regression: first learn a function to approximate \( f_{\text{Reg}}(X, A) \approx E_{\pi^{\text{nat}}}[u|X, A] \) in the observational data; then apply this regression for every individual where \( \pi \) suggests giving the treatment: \( \mathcal{E}_{\text{Reg}}(\pi) = E_{\pi^{\text{nat}}}(X, A)[f_{\text{Reg}}(X, A)\pi(X, A) = 1] \). 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 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. With \( C_i = 1[\pi(X_i, A_i) = T] \), the estimator is

\[
\mathcal{E}_{\text{DR}} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{C_i(\pi)u_i}{P(C_i(\pi) = 1|X_i, A_i)} - \frac{C_i(\pi) - P(C_i(\pi) = 1|X_i, A_i)}{P(C_i(\pi) = 1|X_i, A_i)} f_{\text{Reg}}(X_i, A_i) \right].
\]

We can use an analogous estimator for \( \Delta \), where the same...
We generate observational data from the SCM in Figure 4a, \( V \). The overall objective is varying values of these thresholds. Figure 5 shows that the \( \Delta \) data. The x-axis represents single threshold policies. Estimation \( E \) bly robust) for off-policy estimation of \( E \). Figure 5: Comparing error of \( \pi \) off-policy estimate of \( \Delta \) policy with the highest off-policy estimate of the objective \( \pi \) achieves lower off-policy estimation error on both sensitive groups, across the threshold range. Note the high estimation error of the baseline \( E \). We compute the estimators \( \mathcal{E}_{\text{Reg}}(\pi_j) \) and \( \mathcal{E}_{\text{DR}}(\pi_j) \) for varying values of these thresholds. Figure 5 shows that the causally motivated estimator \( \mathcal{E}_{\text{DR}} \) achieves lower off-policy estimation error on both sensitive groups, across the threshold range. Note the high estimation error of the baseline \( \mathcal{E}_{\text{Reg}} \) for low values of \( \tau \). 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 \( \mathcal{E}_{\text{EqOpp}} \). The overall objective is \( \mathcal{V}_{\pi} = \mathcal{U} - \lambda \mathcal{E}_{\text{EqOpp}} \). We hope to maximize this, with some hyperparameter \( \lambda \in \mathbb{R} \) governing the tradeoff. We note that estimating \( \mathcal{E}_{\text{EqOpp}} \) 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 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 \( \mathcal{E}_{\text{DR}} \) 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 \( X \) 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.
Figure 7: Policy evaluation under credit bureau intervention $\hat{f}_X(X) = \min(X, \tau_{CB})$ with $\tau_{CB} = 600$. Group score change—formally $E_{P \sim do(\tau_X \rightarrow X, f_T \rightarrow Y)}[\Delta_i]$—and institutional profits—$E_{P \sim do(\tau_X \rightarrow X, f_T \rightarrow Y)}[U]$—are shown as functions of the two group thresholds $\{\tau_i\}$ under several fairness constraints.

\[ \hat{X}_i = f_X(X_i) \] as a score that is deterministically computed by the agency from the features (see Fig. 4b). When $f_X$ is the identity function, we recover the original model. Policy evaluation under double intervention $\mathcal{M}_{do(f_T \rightarrow f_T, f_X \rightarrow f_X)}$ 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 $E[U]$ and average per-group score change $E[\Delta_i]$ 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 $f_X(X) = \min(X, 600)$. 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 unlikely to repay, and disproportionately belong to the minority group).

### 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.\(^\text{13}\)

**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 struct-

\(^{13}\)“Mismatch” refers here to structural equations with misspecified functional forms, not incorrect assumptions of causal structure.
Causal Modeling for Fairness in Dynamical Systems

Figure 9: Evaluating multi-step policy robustness to distribution shift for various choice of intervention distribution $q$. Sensitivity of institutional profit—and sensitivity of group avg. score change—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.

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 outcomes are particularly sensitive to assumptions around outcome distributions, policies that capture optimal rewards across many interventional settings are more sensitive to these interventions.

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


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