Inferring Rewards from Discrimination

Team={David Reber}

1. Personal Background

1.1. Relative Academic Strengths

Relative to the class, I claim a strong and diverse *theoretical* background. Above-average areas, with decreasing *relative* expertise:

- **Mathematical Analysis:** *4 undergraduate courses, 2 graduate courses, master's thesis.* Linear/Nonlinear analysis, metric spaces, complex analysis, spectral theory/calculus, measure theory; manifolds, contraction mappings, Frechet derivative, calculus of variations.
- **Linear Algebra:** *3 undergaduate courses, 1 graduate course, master's thesis*. Linear operators, Banach spaces, all canonical decompositions, pseudospectra/pertubation theory; pseudo-inverses, Perron-Frobenius, iterative methods.
- **Dynamical Systems (discrete):** *1 undergraduate courses, 2 graduate courses, master's thesis.* Linear bounding of nonlinear asymptotics, spectral properties of discrete systems, time-varying systems, time delays; isomorphisms, bifucations, chaos theory.
- **Algorithms:** *3 undergraduate courses, 1 graduate, and I led a competitive coding seminar.* complexity, data structures, theory of computation, recursion, random algorithms.
- **Optimization:** *2 undergraduate courses.* unconstrained, linear, convex, nonlinear constrained, dynamic, combinatorial; KKT, weak/strong duality, Monte Carlo.
- **Machine Learning (theory):** *3 undergraduate courses, 2 graduate courses, 2 years industry.* Kernel-based methods, bayesian methods, stochastic dynamic optimization/reinforcement learning, deep learning, statistical theory (measure theory, Bayesian statistics); value iteration, bandit problems.

1.2. Non-academic Expertise

Additionally, I have been interested in anything that can be used to predict agent behavior prior to model training, and causal-based **agent incentives** seem like a great approach. I've read 3 papers on the topic.

I am well-versed with the latest **AI safety** research and proposed agendas from academic institutes such as CHAI (Berkeley) and FHI (Oxford), as well as Deepmind, OpenAI, Anthropic, Redwood Research, and MIRI, and I'm becoming increasingly connected with researchers in the field.

I enjoy **decision theory**; I've read several papers and am well-versed with the variations of Newcomb's paradox and the Prisoner's dillema, Löb's theorem, and the limitations of CDT and EDT generally. *I would love exploring the cooperative game-theoretic implications of Regret Decision Theory (RDT), theoretically or empirically*.

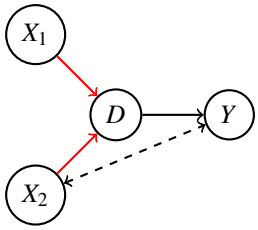
2. Topic and Problem Summary

This proposal best aligns with the **Fairness-Discrimination** bucket, and is adjacent to the **Causal Reinforcement Learning** bucket (see Section 5 for discussion).

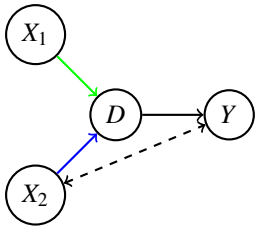
2.1. Motivation and Problem Statement

As algorithms continue to automate systemic decisions (as already attempted with predicted recidivism, loan approval, recruitment/hiring, and academic admissions), there is an increasing need to ensure the fairness of these algorithms. To facilitate better engineering and proactive regulation, we need answers to two questions: 1. What does fairness mean, and 2. How can unfairness be avoided?

R-30 [1] helps answer the first question, by decomposing the total variation into three distinct measures of counterfactual fairness: $Cft\text{-}DE$, $Cft\text{-}IE$, and $Cft\text{-}SE$. However, these measures are non-identifiable from observational data in many non-markovian settings (i.e. counfounding between a mediator and the decision). Furthermore, while these measures point out where the problem lies with surgical precision, they don't really explain *why* the discrimination occured, and consequently, how to promote future fairness (short of blinding the decision-maker to the sensitive paths, which may not be possible).



(a) Direct effect of $\mathbf{X} = \{X_1, X_2\}$ on D



(b) Decomposition of the direct effect of $\mathbf{X}$ on D

Figure 1: A simple example of an incentives decomposition of the direct effect of $\mathbf{X}$ on D, as rewarded by Y.

By using response incentives [3, 4] to augment $Cft\text{-}DE$, $Cft\text{-}IE$, and $Cft\text{-}SE$ (here called Cft quantites), this proposal aims to

- Identify incentivized $Cft^{\uparrow}$ quantites from observational distributions, on causal diagrams where (plain) Cft quantites are non-ID.
- Demonstrate that $Cft^{\uparrow}$ and $Cft^{\downarrow}$ provide a legitimate decomposition of a Cft measure, so that certain types of reward replacement (e.g. through automation) can be predicted to set $Cft^{\downarrow} = 0$, reducing the overall Cft quantity of interest.
- Provide a prodecure for evidencing the reward-dependancies which influenced the decision-maker, to better inform blame attribution and policy adjustment.

Note that the final goal is to determine policy and/or reward changes for an agent or institution, based on observed discrimination and causal knowledge (hence the connection to CRL). Concrete problem statements are included within the research plan of Section 4.

3. Background: Expressing Reward Incentives in an SCM-framework

The following notation and definitions help keep consistency with the lab's existing work on both fairness [1] and CRL [2] (see Figure 1a for a visual aid). In addition to the usual SCM $M = (V, U, F, P(U))$, we use Y as the reward variable(s), $D \in An(Y)$ as the decision node, and $X \in An(D)$ as the sensitive attribute(s) which the decision may be unfair towards. $\mathbf{C} \subset V$ are the covariates which D is allowed to observe; that is, $Pa(D) \subset \mathbf{C}$. A policy π is a soft intervention on D which respects $domain(\pi) \subset \mathbf{C} \cup \{U_D\}$, and an *optimal policy* is defined as a policy π that maximizes the sum of the expected rewards: $\mathbb{E}_{\pi}[\sum_{Y \in \mathbf{Y}} Y]$. Lastly, a 'fairness effect E' is just some element $E \in \{Cft\text{-}DE,\ \ Cft\text{-}IE,\ \ Cft\text{-}SE\}$.

Definition 1 (Response Incentive) *Let $M = (V, U, F, P(U))$ be an SCM. A policy π responds to a variable $X \in V$ if there exists some intervention $do(X = x)$ and some exogenous unit $U = u$, such that $D_x(u) \neq D(u)$. The variable X has an* response incentive *if all optimal policies respond to X. We say a causal diagram G* admits *an response incentive on X if it is compatible with an SCM that has an response incentive on X.*

Note that *response incentive* is a binary classification of the ancestors of D: either a node has an response incentive, or it doesn't. Next, the *minimal reduction* of a causal graph G eliminates all edges from parents of D which do not have their own d-connection to Y.

Definition 2 (Minimal Reduction) *The* minimal reduction G^{min} *of a causal diagram G is the result of removing from G all edges $X \to D$ satisfying $(X \perp\!\!\!\perp Y | D \cup Pa(D) \setminus X)$.*

Intuitively, the minimal reduction breaks all non-informative links to the decision node. The main result from [3] we build on is a graphical criterion for classifying a node X as having an response incentive.

Theorem 1 (Markovian Case: Response Incentive Criterion [3]) *Let G be a Markovian causal diagram. Then G admits a response incentive on $X \in V$ if and only if the minimal reduction G^{min} has a directed path X - $\to D$.*

Definition 3 (Incentivized Discrimination) *Let G be a causal diagram, with X, D, and Y as previously defined. Let E be some measure of a discriminatory effect X on D. The* incentivized effect $E^{\uparrow}$*, of X on D with respect to Y, is defined as the value of E on G^{min}.*

For example, consider the counterfactual direct effect $Cft\text{-}DE(X, D)$ of X on D. Then $Cft\text{-}DE^{\uparrow}(X, D, Y)$ is obtained by first finding G^{min} (which is defined only relative to Y), then computing the $Cft\text{-}DE$ effect in this reduced graph (see Figure 2).

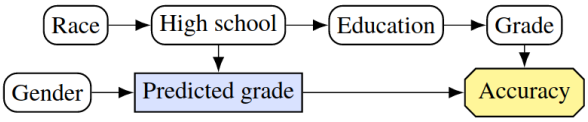


Figure 2: Source: [3]. Here, discrimination is incentivized with respect to Race, but not incentivized with respect to Gender, because the edge from Gender is eliminated in G^{min}.

4. Research Plan and Preliminary Work

The first step will be to extend Theorem 1 to the Non-Markovian case: this is foundational for the following directions. (It seems trivial if confounding is not allowed on D, but then we wouldn't be able to say anything about the Berkeley admissions case).

4.1. Identifying Incentivized Discrimination

The identification problem of the Cft quantities is strictly easier when incentives are taken into account, since the reduced graph G^{min} is obtained by removing edges from G. I conjecture there will be nontrivial families of graphs for which $Cft\text{-}DE$, $Cft\text{-}IE$, and $Cft\text{-}SE$ are not identifiable from observational distributions, but for which $Cft\text{-}DE^{\uparrow}$, $Cft\text{-}IE^{\uparrow}$, and/or $Cft\text{-}SE^{\uparrow}$ are always identifiable.

Problem 1 *Identification of Incentifized Cft quantities*
Inputs: *G, $P(V)$, and an incentivised effect $E^{\uparrow} \in \{Cft\text{-}DE^{\uparrow},\ Cft\text{-}IE^{\uparrow},\ Cft\text{-}SE^{\uparrow}\}$*
Outputs: *Yes/No (Is $E^{\uparrow}$ ID from G and $P(V)$)*

4.2. Incentives-based Decomposition of Discrimatory Effects

Conjecture 1 (Unincentivized Null) *Let G be a causal diagram, with X, D, and Y as previously defined. Suppose $(X \perp\!\!\!\perp Y | D \cup Pa(D) \setminus X)$.*
Then $Cft\text{-}DE^{\uparrow}(X,D) = 0$.
(Epistemic status: Almost certain. I speculate that there may be similar results about $Cft\text{-}IE^{\uparrow}(X,D)$ and $Cft\text{-}SE^{\uparrow}(X,D)$, but I'm less confident about those.)

If our models and data were perfect, and our policies optimal in maximizing rewards, then (assuming Conjecture 1 holds) we will observe that $Cft\text{-}DE^{\uparrow}(X,D) = 0$. But what if our model is mispecified, our data is noisy, or our policies are sub-optimal for maximizing our reward Y? Then our null result won't be exactly zero; perhaps only close to zero.

How might we measure the magnitude of this 'unincentivized' effect from data, when theoretically it should be zero? Here's a possible working definition:

Working Definition 1 (Unincentivized Discrimination) *Let G be a causal diagram, with $\mathbf{X}$ containing multiple sensitive attributes. Partition $\mathbf{X} = \mathbf{X}^{\uparrow} \cup \mathbf{X}^{\downarrow}$, where $\mathbf{X}^{\downarrow} := \{X \in An(D) | (X \perp\!\!\!\perp Y | D \cup Pa(D) \setminus X)\}$ and $\mathbf{X}^{\uparrow} := \mathbf{X} \setminus \mathbf{X}^{\downarrow}$.*
Let $E(X,D)$ be some fairness measure of a discriminatory effect X on D. The unincentivized effect *$E^{\downarrow}(\mathbf{X},D,Y)$, of X on D with respect to Y, is defined as*

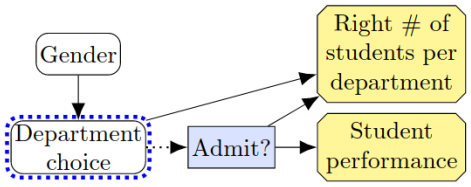
$$E^{\downarrow}(\mathbf{X},D,Y) := E(\mathbf{X}^{\downarrow},D)$$

Note that Conjecture 1 could now be expressed as saying that $Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y) = 0$ (if our models and data are perfect and our policies are optimal in maximizing rewards).

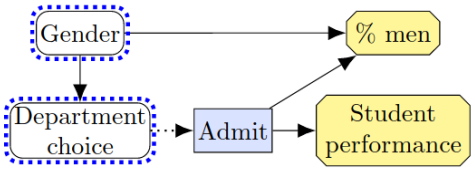
No matter what definition of 'Unincentived Discrimination' we use, we want to ensure that it's physically (and intuitively) meaningful; that is, the 'incentivized' and 'unincentivized' portions of the $Cft\text{-}DE$ effect should form a **decomposition** of the total value of the $Cft\text{-}DE$ effect.

Problem 2 *Incentivized Decomposition of Discrimination*
Inputs: *An incentivised effect $E \in \{Cft\text{-}DE,\ Cft\text{-}IE,\ Cft\text{-}SE\}$*
Outputs: *A function g satisfying $E(\mathbf{X},D,Y) = g(E^{\uparrow}(\mathbf{X},D,Y), E^{\downarrow}(\mathbf{X},D,Y))$ (which holds across all G and $P(V)$)*

Conjecture 2 (Decomposition of Discrimantory Effect) *$E(\mathbf{X},D,Y) = E^{\uparrow}(\mathbf{X},D,Y) + E^{\downarrow}(\mathbf{X},D,Y)$.*
(Epistemic status: Somewhat confident. I think it's more likely to hold in the linear case, at any rate).



(a) Reward Y_2 which does not incentivize bias



(b) Reward Y_2 which does incentivize gender bias

Figure 3: Source: [4]. Using our notation, $X =$'Gender', $D =$'Admit', $Y_1 =$'Student performance', and either $Y_2 =$'Right number of students' or $Y_2 =$'percent men'. Although the reward is not known to us, we can infer its structure based on how well it matches the discrimination.

4.3. Explaining Discrimination via Rewards

If we have discriminatory data and a causal diagram, we would like to be able to assign likelihoods to the various reward structures which could explain the discrimination.

Problem 3 *Evidence of Reward*

Inputs: *G_Y (G with node Y removed), $P(V)$, $Pa(\hat{Y})$ (a candidate set of Y's parents), and a value for $E(X,D)$ (where $E \in \{Cft\text{-}DE,\ Cft\text{-}IE,\ Cft\text{-}SE\}$)*

Outputs: *Fraction of $E(X,D)$ explainable by $Pa(\hat{Y})$*

Note that for a given dataset and casual diagram G, the value of $Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y)$ is dependent on the structural assumptions of Y: in particular, on the dependencies that Y has.

Suppose that, for proprietary or historical reasons, we do not know Y's dependancies (as in Figure 3). In this case, perhaps we can use $Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y)$ as evidence for whether our modeling choice of Y is valid, since if $Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y) < Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y')$, then presumably the reward Y is explaining more of the discrimination than Y' is. Formally:

Conjecture 3 (Reward Identification) *Let G, $P(V)$ be given, with $\mathbf{X}$ and D specified. Assume $P(V)$ was generated by an optimal policy π^* w.rt. the true, unknown reward function f_{Y^*}. Then $Pa(\hat{Y}) \subset Pa(Y^*)$, where $\hat{Y}$ satisfies*

$$\hat{Y} = \arg\min_{Y} \quad Cft\text{-}DE^{\downarrow}(\mathbf{X},D,Y)$$

If f_{Y^*} has no trivial dependancies, then $Pa(Y^*) = Pa(\hat{Y})$

(Epistemic status: Likely. I suspect 'trivial dependencies' are important to consider, but I recognize that's a hand-wavy term. I'm trying to convey that Y^* and $\hat{Y}$ share the same dependencies, and there may be ways to express that other than $Pa(\hat{Y}) = Pa(Y^*)$).

I would like to test Conjecture 3 numerically using synthetic datasets first, to first get a sense for its usefulness (since it seems likely that noisy data, or a suboptimal policy, might render such a conjecture useless in a practical sense).

5. Relation to CRL

Ultimately, the goal of this research direction is to predict fairness attributes of an ML architecture prior to training, to make 'being fair' easier to engineer, and easier to regulate. But, this is only possible given some degree of causal knowledge about the feature space (and the reward function, but that's engineered).

I expect these results to evidence that CRL is inherently easier to align with what we want (such as 'fairness') than non-causal RL. However, it is not immediately clear to me how addressing any of these problems can improve performance/training, unfortunately.

I was surprised to realize that Problem 3 is very similar to causal imitation learning: the difference is that instead of trying to learn an agent's policy, we are trying to infer some information about about agent's reward structure.

Meanwhile, addressing Problem 1 *may* enable $Cft\text{-}DE$, $Cft\text{-}IE$ and $Cft\text{-}SE$ to be used more readily as regularization functions.

References

1. J. Zhang, E. Bareinboim, "Fairness in Decision-Making – The Causal Explanation Formula", AAAI-18. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018. Columbia CausalAI Laboratory, Technical Report (R-30), Nov, 2017. https://causalai.net/r30.pdf
2. J. Zhang, E. Bareinboim, "Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach", ICML-20. In Proceedings of the 37th International Conference on Machine Learning, 2020. Columbia CausalAI Laboratory, Technical Report (R-57), Jun, 2020. https://causalai.net/r57.pdf
3. T. Everitt, R. Carey, E. D. Langlois, P. A. Ortega, S. Legg, "Agent Incentives: A Causal Perspective," in *Proceedings of the AAAI 2021 Conference*, arXiv:2102.01685v2 [cs.AI], 2021. https://arxiv.org/abs/2102.01685v2
4. T. Everitt, P. A. Ortega, E. Barnes, S. Legg, "Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings," arXiv:1902.09980v7 [cs.AI], 2022. https://arxiv.org/abs/1902.09980v7