A New Paradigm for Counterfactual Reasoning in Fairness and Recourse

Lucius E.J. Bynum[1] , **Joshua R. Loftus**[2] and **Julia Stoyanovich**[1,3]
[1]New York University, Center for Data Science
[2]London School of Economics, Department of Statistics
[3]New York University, Tandon School of Engineering
lucius@nyu.edu, J.R.Loftus@lse.ac.uk, stoyanovich@nyu.edu

Abstract

Counterfactuals underpin numerous techniques for auditing and understanding artificial intelligence (AI) systems. The traditional paradigm for counterfactual reasoning in this literature is the *interventional counterfactual*, where hypothetical interventions are imagined and simulated. For this reason, the starting point for causal reasoning about legal protections and demographic data in AI is an imagined intervention on a legally-protected characteristic, such as ethnicity, race, gender, disability, age, etc. We ask, for example, *what would have happened had your race been different?* An inherent limitation of this paradigm is that some demographic interventions — like interventions on race — may not be well-defined or translate into the formalisms of interventional counterfactuals. In this work, we explore a new paradigm based instead on the *backtracking counterfactual*, where rather than imagine hypothetical interventions on legally-protected characteristics, we imagine *alternate initial conditions* while holding these characteristics fixed. We ask instead, *what would explain a counterfactual outcome for you as you actually are or could be?* This alternate framework allows us to address many of the same social concerns, but to do so while asking fundamentally different questions that do not rely on demographic interventions.

1 Introduction

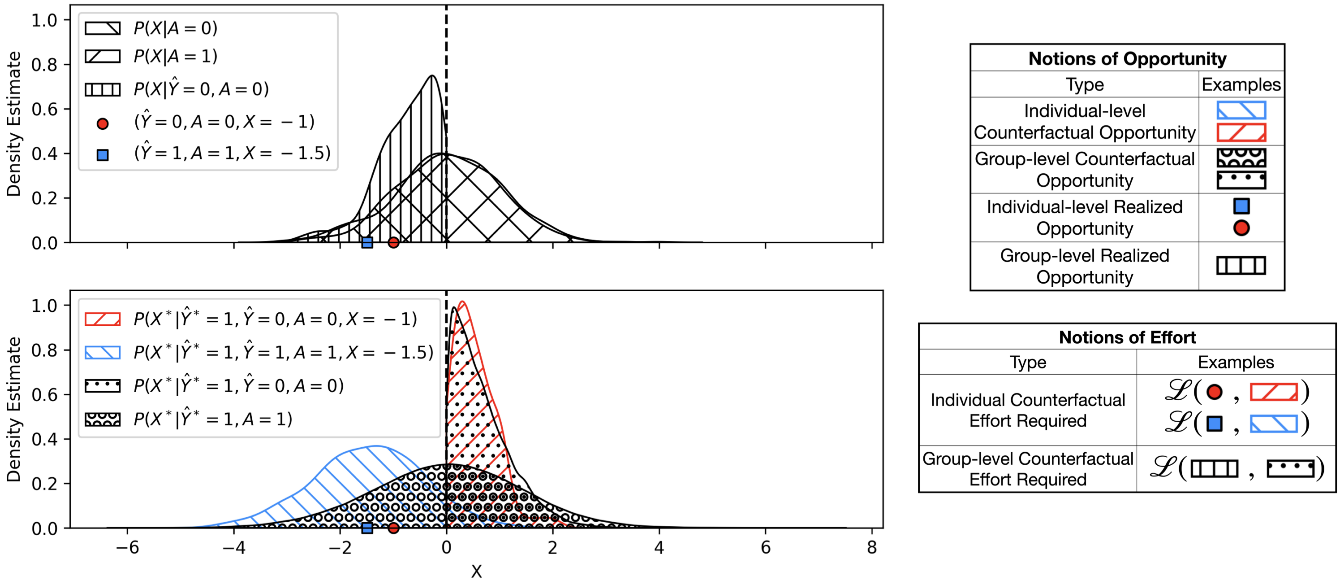
Counterfactual reasoning plays a pivotal role in many modern techniques for auditing and understanding machine learning models and artificial intelligence (AI) systems. Traditionally, counterfactual discrimination criteria ask the following question: *How would this decision-making system behave if, counterfactually, an individual had a different value of their legally-protected characteristic?* For example, would a given person have more likely received a financial loan if they were not in a minority group? Depending on the setting, such behavior from a decision-making system can be seen as evidence of discrimination [Kilbertus *et al.*, 2017; Nabi and Shpitser, 2017; Loftus *et al.*, 2018].

While causal reasoning, and counterfactual reasoning in particular, open important avenues of analysis, the primary starting point for these analyses — an imagined intervention on a legally-protected characteristic — is also a limitation. We summarize two key reasons for this limitation. One reason is that legally-protected characteristics are not always defined in a way that is compatible with interventional counterfactuals. Consider, for example, the social category *race*: it is often unclear what intervention on someone's race would actually correspond to and remains an open question whether or not such interventions can be well-defined [Benthall and Haynes, 2019; Hanna *et al.*, 2020; Kasirzadeh and Smart, 2021; Sen and Wasow, 2016; Jacobs and Wallach, 2021]. A second reason for this limitation is that social categories are at times intertwined with other features of interest. Think, for example, about analyzing US Census data, which has features like birthplace, racial category, socioeconomic status, and ethnicity. These features are often used to *define* each other, and thus cannot easily be separated into distinct interventions.[1] However, reasoning about social categories like race is crucial if we wish to account for real-world disparities. If someone both (1) cares about making more equitable decisions, and (2) believes that causal structure plays an important role in governing the context around a decision-making system, they are left with few tools for counterfactual reasoning that do not suffer from this limitation.

In this paper, we propose a new paradigm for counterfactual reasoning in algorithmic fairness and recourse that alleviates this problem. This paradigm starts with a different counterfactual question: *What would explain a counterfactual outcome for an individual as they actually are or could be?* Starting with this question allows us to still reason counterfactually about fairness and recourse, but to do so without relying on hypothetical demographic interventions.

Underpinning these two different counterfactual questions are two different semantics for computing counterfactuals. The *interventional counterfactual*, where hypothetical interventions are imagined and simulated, is the dominant semantics used for computing counterfactuals across computer science literature. In this context, counterfactual discrimination criteria have primarily been formalized by imagining hypothetical interventions on demographic variables or

[1] We can state this formally in terms of *modularity* (see § 2).



Notions of Opportunity

Type	Examples
Individual-level Counterfactual Opportunity	▨ ▧
Group-level Counterfactual Opportunity	▩ ⋯
Individual-level Realized Opportunity	■ ●
Group-level Realized Opportunity	▥

Notions of Effort

Type	Examples
Individual Counterfactual Effort Required	$\mathscr{L}$(● , ▧) $\mathscr{L}$(■ , ▨)
Group-level Counterfactual Effort Required	$\mathscr{L}$(▥ , ⋯)

Figure 1: A visual example of several backtracking counterfactual quantities to consider for algorithmic fairness and recourse (labeled on the right). Based on Example 1, consider covariate $X \sim \mathcal{N}(0, 1)$, protected attribute $A \sim$ Bern(0.5), and predictor $\widehat{Y} = A \vee (X > 0)$. A corresponding DAG is shown in Figure 2b. With backtracking conditional $P_B(U^* \mid U) = (U_A^* = U_A, U_X^* = \mathcal{N}(U_X, 1), U_Y^* = 0)$ and an appropriate cost function $\mathcal{L}$, we can use Algorithm 1 to estimate the various notions of opportunity and effort shown above for both individuals and groups. This process is formally defined in Sections 3 & 4.

their proxies [Kusner *et al.*, 2017; Chiappa, 2018; Wu *et al.*, 2019; Plecko and Bareinboim, 2022; Kilbertus *et al.*, 2017; Zhang and Bareinboim, 2018; Von Kügelgen *et al.*, 2022; Ehyaei *et al.*, 2023; Carey and Wu, 2022]. Recent work has formalized an alternate semantics for computing counterfactuals: the *backtracking counterfactual* answers counterfactual queries by imagining changes to initial conditions instead of simulating hypothetical interventions [Von Kügelgen *et al.*, 2023]. Just as the interventional counterfactual is a natural fit for reasoning about demographic interventions, the backtracking counterfactual is a natural fit for explaining counterfactual outcomes, and thus, a natural fit for our alternate counterfactual question.

Contribution. In what follows, we define new technical notions of counterfactual opportunity and effort, and introduce several new counterfactual discrimination criteria. These are, to our knowledge, the first counterfactual-based fairness criteria that depart from the need to consider intervention on legally-protected characteristics, while still allowing us to simultaneously (1) consider legally-protected characteristics, (2) make use of causal relationships between variables, and (3) consider fairness at an individual or a group level. We also develop a simple algorithm for sampling backtracking counterfactuals and use it to implement several of the proposed criteria on real and simulated data. In effect, this paper lays the foundation for a potentially large body of work. We view this as the first step in a wider research agenda that charts a course for future work across algorithmic fairness and recourse and, more broadly, develops a new approach to counterfactual analysis for demographic data.

The core idea behind the counterfactual discrimination criteria we introduce in this work is to quantify unfairness for an individual or group in terms of their *opportunity* or their *required effort* — factually and counterfactually. Using these criteria, we can answer the following types of questions:

- What opportunities, factually and counterfactually, are available to different individuals?
- Does access to opportunity track group membership?
- How much effort would be required to get a different outcome, both for individuals and across groups?

As a running example, consider an algorithm that is used to assist in hiring decisions or recruitment.

Example 1 (Discrimination in Hiring). *Let binary variable A represent whether or not a person is in a majority group, binary Q represent whether or not some job-related qualifications X exceed a desired threshold, and binary $\widehat{Y}$ represent whether or not they receive a job offer. Consider the following discriminatory hiring practice: $\widehat{Y} = A \vee Q$, or in other words, a person is offered the job either if they are in the majority group or if they have the relevant qualifications.*

At the level of intuition, in Example 1 we can say that a person in the minority group with $A = 0$ has only one opportunity to get the job — they need $Q = 1$ — whereas people in the majority group with $A = 1$ have more opportunity to get the job, including both $Q = 0$ and $Q = 1$. [2] See Figure 1

[2] For the reader concerned about 'fairness through blindness,' the operative question is this: *Is the variable Q measuring only morally relevant qualifications, or is it also measuring different levels of access to resources?*

for a visualization of several ways we can use backtracking counterfactuals to represent opportunity and effort in Example 1. In the remainder of this paper, we formalize how to compute the quantities in Figure 1, both for this example and for arbitrary settings with possibly larger sets of variables and more complicated relationships between them.

2 Mathematical Preliminaries

In this section, we provide the necessary mathematical background to define our counterfactual discrimination criteria, with structural causal models and interventional counterfactuals following [Pearl, 2000] and backtracking counterfactual semantics and notation from [Von Kügelgen *et al.*, 2023].

Causal models. We define a *causal model* as a tuple $\mathcal{M} = (V, U, F)$. In this tuple, V is a set of observed variables, U a set of unobserved (exogenous) variables, and F a set of functions $\{f_i\}_{i=1}^{|V|}$ for each $V_i \in V$ such that $V_i = f_i(P_i, U_{P_i})$ where $P_i \subseteq V \setminus \{V_i\}$ represents the causal parents of V_i and $U_{P_i} \subseteq U$. A causal model can be pictorially represented as a directed acyclic graph (DAG) with nodes for U, V and directed edges for F. A *probabilistic causal model* $(\mathcal{M}, P(U))$ adds distribution $P(U)$ over the unobserved variables.

Interventions and counterfactuals. We define an atomic *intervention* on variable V_i as the substitution of equation $V_i = f_i(P_i, U_{P_i})$ with a particular value $V_i = v$, simulating the forced setting of V_i to v by an external agent, commonly notated do$(V_i = v)$. A *submodel* $\mathcal{M}_x$ for realization x of $X \subseteq V$ is the model $\mathcal{M}$ after intervention do$(X = x)$. Given a probabilistic causal model, we can derive from $(\mathcal{M}_x, P(U))$ the distribution of any subset of variables following intervention do$(X = x)$. Using instead a distribution over the exogenous variables that is specific to a particular context or individual, the same mechanics allow us to define *interventional counterfactuals* to model alternate possible outcomes after interventions in a specific context. Following [Von Kügelgen *et al.*, 2023; Balke and Pearl, 1994], we use an asterisk to denote counterfactual versions V^* of variables V. Counterfactual variable Y^* given a factual observation z and intervention do$(X = x^*)$ $(X, Y, Z \subseteq V)$ can be computed via a three-step procedure often referred to as 'abduction, action, prediction.' Abduction uses observed evidence to obtain $P(U \mid z)$ from $P(U)$. Action performs intervention do$(X = x^*)$ for counterfactual realization x^* to obtain $\mathcal{M}_{x^*}$. Prediction computes the probability of Y^* from $(\mathcal{M}_{x^*}, P(U \mid z))$.

Backtracking counterfactuals. Backtracking counterfactuals suggest instead a semantics of counterfactuals that *leaves all of the causal mechanisms unchanged.* Instead of external actions being simulated via intervention, alternate possible outcomes are explained via changes to the values of exogenous variables. With unchanged mechanisms, there are many possible exogenous changes that can explain counterfactual observations. Backtracking counterfactuals answer probabilistic queries about these changes by first specifying the *backtracking conditional* $P_B(U^* \mid U)$, a similarity measure between factual exogenous conditions U and counterfactual exogenous conditions U^*. This, in turn, induces a joint distribution $P_B(U^*, U) = P_B(U^* \mid U)P(U)$ that allows for the computation of counterfactuals through a 'cross-world abduction, marginalization, and prediction' procedure similar to abduction, action, and prediction [Von Kügelgen *et al.*, 2023]. Consider again counterfactual variable Y^* given a factual observation z and — this time, instead of an intervention — *counterfactual observation* $X = x^*$. In cross-world abduction, $P_B(U^*, U)$ is updated with evidence (x^*, z) to obtain $P_B(U^*, U \mid x^*, z)$. In marginalization, U is marginalized out to obtain $P_B(U^* \mid x^*, z)$. Finally, prediction computes the probability of Y^* from model $(\mathcal{M}, P_B(U^* \mid x^*, z))$. Observe here that no changes to any causal mechanisms have been made — we answer the counterfactual query with $\mathcal{M}$ rather than $\mathcal{M}_{x^*}$. Figure 2 shows a visual representation of this difference for Example 1, showing the DAGs for an interventional (2a) versus backtracking (2b) approach, and the corresponding exogenous conditions.

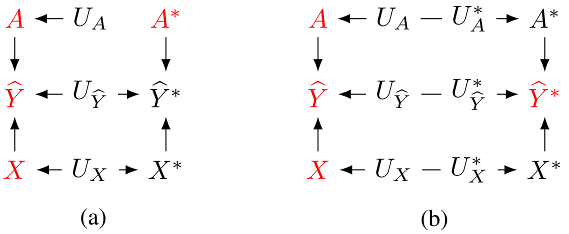
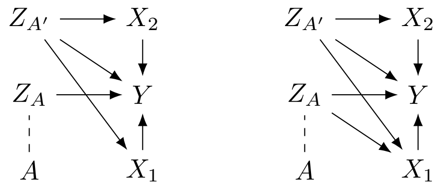


Figure 2: Graphical models for two counterfactual fairness problems with protected attribute A, features X, and classification outcome $\widehat{Y}$. (a) Interventional approach with intervention do$(A = A^*)$. (b) Backtracking approach with counterfactual observation $\widehat{Y}^*$. Red nodes are known. Based on Figure 2 in [Von Kügelgen *et al.*, 2023].

Modularity and social categories. The question of modularity has roots in the philosophy of causal inference and in theories of how cause-and-effect is defined, often discussed in connection with concepts like manipulation [Holland, 1986] or ontological stability [Barocas *et al.*, 2023]. We call a variable in a causal model *modular*, or say that it satisfies *modularity*, if (1) its mechanism remains invariant when other mechanisms are subjected to external influences, and (2) if other mechanisms remain invariant when its mechanism is subjected to external influences [Pearl, 2000, p. 60]. The notion of modularity arises in interventional counterfactual queries through the fact that intervention do$(X = x^*)$ leaves all mechanisms other than $\{f_i : V_i \in X\}$ untouched. However, modularity need not be assumed to answer backtracking counterfactual queries because there are no external influences to any mechanisms (no interventions).

Questions about modularity also arise naturally in discussions of how to conceptualize counterfactual discrimination criteria and how to incorporate social categories in causal models. In this work, we use the term *social category* to refer to a grouping of people, agnostic to whether or not there are any shared attributes within the group. This makes our framework compatible with groupings that are socially constructed — by which we mean, created by people in a society. Questioning whether or not variables like social categories can truly be modular (or intervened on at all) has led to questions



(a) (b)

Figure 3: (a) Stylized causal model for a hiring example with age A, success measure Y, latent variables $Z_{A'} \perp A$ and $Z_A \not\perp A$, and qualifications X_1, X_2. (b) An alternate version of (a) where latent variable Z_A also impacts X_1.

in the research community of whether causal fairness methods, especially with social categories, are valid or reliable [Hu and Kohler-Hausmann, 2020; Kasirzadeh and Smart, 2021; Carey and Wu, 2022]. Past work has contended with this issue for treatment choice and policy learning [Bynum *et al.*, 2021; Bynum *et al.*, 2023]. Notably, in this work, the semantics of backtracking counterfactuals allows us to avoid assuming modularity. Our framework thus provides *a possible answer to how we can relax the modularity assumption* and still conceptualize counterfactual discrimination criteria.

3 Estimating Opportunity

In this section, we detail how backtracking counterfactuals can serve as a useful formalism to describe *opportunity* in the context of an algorithmic decision.

A backtracking conditional to capture agency. Recall the backtracking conditional $P_B(U^* \mid U)$ serves as a similarity measure between factual conditions U and counterfactual conditions U^*. In other words, $P_B(U^* \mid U)$ tells us which counterfactual worlds to consider (and how heavily to weight them) in an answer to a counterfactual query. In a setting where the agency of an individual is at play, rather than use counterfactuals to estimate what changes are *probable* for an individual to get a different outcome, a more natural first step is to consider what changes are *possible* and then, separately, how difficult those changes are to achieve. For this reason, we propose the following 'non-informative' backtracking conditional that focuses on what an individual would or would not be able to change (their mutable vs. immutable variables).

Definition 1 (Non-informative Backtracking Conditional Distribution)**.** *Consider a partitioning of the variables U into mutable variables M and immutable variables $U \setminus M$. We define a non-informative backtracking conditional distribution as $P_B(U^* \mid U = (m, n)) = P_M(m)\delta(n)$. Here, $P_M(m)$ is the prior marginal distribution over the variables $m \in M$ and $\delta(n)$ is a Dirac delta distribution (i.e., a point mass) at $n \in U \setminus M$.*[3]

The non-informative backtracking conditional makes sense if an individual is capable of changing their mutable covariates an arbitrary amount in order to achieve the goal of changing their outcome. But, to capture other worldviews on agency or encode additional constraints on how much an individual could change their covariates, an arbitrary backtracking conditional $P_B(U^* \mid U)$ can be used, under the restriction that we only consider the conditional $P_B(U^* \mid M = m, N = n, N^* = n)$ in our analysis. In both cases, the underlying concept is that an individual can change only their mutable characteristics. With a means of capturing agency, we now define opportunity through the concept of an *opportunity set*.

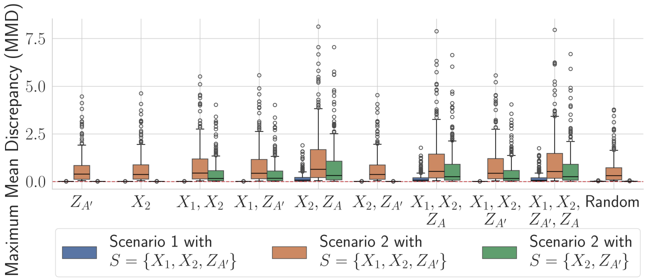


Figure 4: Visualization of how well Individual Equality of Counterfactual Opportunity is satisfied for models using different subsets of covariates across Scenarios 1 and 2, calculated as the absolute energy distance MMD between the two terms of Definition 7.

Definition 2 (Opportunity Set)**.** *Given causal model $\mathcal{M} = (V, U, F)$, an opportunity set $S \subseteq U \cup V$ is a set of mutable covariates that we believe capture opportunity in the context of a particular algorithmic decision.*

Returning to Example 1, we could set $S = \{X\}$ to capture that job-related qualifications X are the mutable covariates that represent an individual's opportunity to get $\widehat{Y} = 0$ or $\widehat{Y} = 1$. In general, given an opportunity set S and a backtracking conditional distribution that describes agency, we can estimate what opportunities different individuals and groups have — both factually and counterfactually — to receive particular algorithmic decisions. Across Definitions 3-8, let A, X, Y represent the protected attributes, remaining attributes, and output of interest, respectively; assume probabilistic causal model $(\mathcal{M} = (V, U, F), P_U)$, where $V \equiv A \cup X$; and consider opportunity set $S \subseteq U \cup V$.

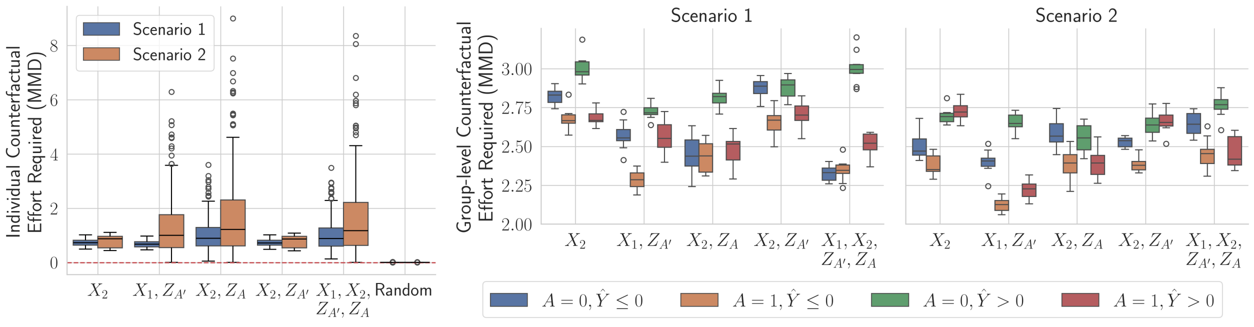
Definition 3 (Individual-level Counterfactual Opportunity)**.** *Consider individual observation $X = x, A = a, \widehat{Y} = y$. The opportunity that this individual has to achieve a counterfactual outcome $\widehat{Y}^* = y^*$ is given by the distribution $P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y, X = x, A = a)$.*

Definition 4 (Individual-level Realized Opportunity)**.** *Consider individual observation $X = x, A = a, \widehat{Y} = y$. The realized opportunity this individual had to achieve outcome y is given by their observed value s of opportunity set S.*

Definition 5 (Group-level Counterfactual Opportunity)**.** *Consider a group of individuals defined by value g of covariates G.*[4] *The opportunity that this group has to achieve a counterfactual outcome $\widehat{Y}^* = y^*$ is given by the distribution $P(S^* \mid \widehat{Y}^* = y^*, G = g)$.*

[3] This definition implicitly requires a factorizable original distribution $P(U)$, as is typically the case in an SCM.

[4] $G = g$ can be, for example, a setting of protected attributes A, a setting of covariates X, or both.



(a)

(b)

Figure 5: Visualization of (a) Individual Counterfactual Effort Required (Definition 9) and (b) Group-level Counterfactual Effort Required (Definition 10) for models using different subsets of covariates across Scenarios 1 and 2, with energy distance MMD for cost function $\mathcal{L}$.

Definition 6 (Group-level Realized Opportunity). *Consider a group of individuals defined by value g of covariates G. The realized opportunity this group had to achieve outcome y is given by the distribution $P(S \mid \widehat{Y} = y, G = g)$.*

Figure 1 shows visual examples of Definitions 3 - 6 for Example 1. An immediate consequence of these computational descriptions of opportunity is that we can use them to define several new notions of equality of opportunity.

Definition 7 (Individual Equality of Counterfactual Opportunity). *Consider individual observation $X = x, A = a, \widehat{Y} = y$. Predictor $\widehat{Y}$ satisfies individual equality of counterfactual opportunity for this individual if for all y^*,*

$$P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y, X = x, A = a) = P(S^* \mid \widehat{Y}^* = y^*).$$

In words, Definition 7 captures that an individual's counterfactual opportunities to get an outcome should be the same as the overall population's. We can similarly draw this comparison across those who got the same outcome, capturing equality of *recourse opportunity* with $P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y)$ instead of $P(S^* \mid \widehat{Y}^* = y^*)$. We could also, by including $X = x$ and/or $A = a'$ as additional conditions on the right hand side, draw a comparison across only those with the same covariates or draw a comparison across groups. In each case, we consider individuals' counterfactual opportunities to get the different possible outcomes. Just as we can compute and equate opportunity for individuals, we can do so for groups.

Definition 8 (Group-level Equality of Counterfactual Opportunity). *Predictor $\widehat{Y}$ satisfies group-level equality of counterfactual opportunity if for all y^* and groups g, g',*

$$P(S^* \mid \widehat{Y}^* = y^*, G = g) = P(S^* \mid \widehat{Y}^* = y^*, G = g').$$

In words, Definition 8 states that each group overall should have the same opportunities to get each outcome. As with Definition 7, we can add additional conditions to either side of Definition 8 for more fine-grained comparisons across groups, e.g., capturing group-level equality of recourse opportunity by adding $\widehat{Y} = y$ to each side.

4 Describing Effort

In Section 3, we make use of backtracking counterfactuals to capture individuals' *opportunities* to achieve an outcome. In this section, we describe how we can instead capture the *effort required* to achieve an outcome: for example, the cost of recourse. Rather than make assumptions about how much effort an individual may or may not exert in a given situation, we focus instead on how much effort would be required of them, characterized as a difference between two possible states rather than a property of the individual. We can use any appropriate cost function to formalize this difference.

Definition 9 (Individual Counterfactual Effort Required). *Consider individual observation $X = x, A = a, \widehat{Y} = y$ and realized opportunity set $S = s$. The effort required for this individual to achieve counterfactual outcome $\widehat{Y}^* = y^*$, with respect to cost function $\mathcal{L}$, is given by*

$$\mathcal{L}\left(s, P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y, X = x, A = a)\right).$$

Definition 9 captures how costly a change from y to y^* is in terms of a distance between what the covariates of interest $S = s$ currently look like compared to what they could look like counterfactually to get the other outcome. The cost function $\mathcal{L}$ represents whatever problem-specific choices we make to account for effort.

Definition 10 (Group-level Counterfactual Effort Required). *Consider opportunity set S and a group of individuals defined by value g of covariates G. The effort required for this group to achieve counterfactual outcome $\widehat{Y}^* = y^*$ given $\widehat{Y} = y$, with respect to cost function $\mathcal{L}$, is defined as*

$$\textsf{\textit{GCE}}(g) \equiv \mathcal{L}\Big(P(S \mid \widehat{Y} = y, G = g), P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y, G = g)\Big).$$

Definition 10 achieves the same comparison as Definition 9, but now for group g instead of an individual.

As with opportunity, we can equate required effort across individuals or groups as additional counterfactual discrimination criteria. Using the logic of Definition 9 to capture effort at an individual level, we are left with the question —

present for any individual-level notion of fairness — of what constitutes a fair baseline for comparison for this individual's effort. Depending on the problem context, Definition 9 can be used to draw whichever individual-level comparison is desired, and can be computed across individuals as a diagnostic tool to explore which individuals to focus on. At the group level, a natural first comparison to draw is across groups.

Definition 11 (Group-level Equality of Effort)**.** *Predictor $\widehat{Y}$ satisfies group-level equality of effort if $\mathsf{GCE}(g) = \mathsf{GCE}(g')$ for all groups g, g'.*

Several notions of effort and the cost of recourse exist in other literature [Karimi *et al.*, 2022], including work such as [Von Kügelgen *et al.*, 2022] that frames recourse as a fundamentally causal problem. Even within causal fair recourse literature, analogous notions of recourse cost and required effort are expressed in terms of interventional counterfactuals, and fairness is conceptualized — particularly at the individual level — via the traditional paradigm of intervention on A. Also unlike other work in recourse, in our case we consider individuals moving both from negative to positive and from positive to negative predictions, more generally capturing a model's behavior. Definitions 9 - 11 are able to capture classical (asymmetric) recourse if, for example, the cost function $\mathcal{L}$ only accounts for counterfactual changes across the decision boundary from factual outcomes $\widehat{Y} = 0$ to counterfactual $\widehat{Y}^* = 1$ and ignores changes from $\widehat{Y} = 1$ to $\widehat{Y}^* = 0$.

5 Experiments

In this section, we demonstrate our criteria computationally on real and simulated data. Using the framework introduced in Sections 3 and 4, a variety of different counterfactual quantities and discrimination criteria can be computed for a given predictor. Exploring all these possibilities computationally is beyond the scope of this short paper: in this section we focus on Definitions 7, 9, and 10 as representative examples of both opportunity-based and effort-based criteria. Code to reproduce all our experiments is available online.[5]

In order to actually compute backtracking counterfactuals, we take a simple approach to estimating the joint distribution $P(U, V, U^*, V^*)$, detailed in Algorithm 1. This algorithm samples n^* backtracking counterfactuals for each of the n observed units. Although the brute-force approach in Algorithm 1 is in theory suitable for any backtracking counterfactual query that conditions on either observed U, V or finite-domain U^*, V^* (given large enough n and n^*), it is by nature not scalable. However, the development of efficient or query-first algorithms for sampling backtracking counterfactuals is not our focus here, and the naive approach is suitable for each of our experiments.

5.1 Synthetic Hiring Datasets

To explore a more complex graphical structure than Example 1, consider the following setup. Figures 3a and 3b show a causal graphical model that pictorially represents how an applicant's qualifications X_1 and X_2 impact some measure of job success Y. In this example, *age group A* is the protected attribute. Everything unmeasured that we believe impacts X_1, X_2 and Y is latent, and for simplicity, we assume we can partition the latent space into variables $Z_A \not\perp A$, representing age-related (immutable) circumstance, and $Z_{A'} \perp A$, representing (mutable) other factors unrelated to age. In this context, imagine an algorithm used during the hiring process that makes a prediction $\widehat{Y}$ of future success.

Algorithm 1 Backtracking Counterfactual Sampling

```
Input: $P(U^* \mid U)$, $\mathcal{M}$, $\{V_i\}_{i=1}^n$
Parameter: $n^*$
Output: $\{(U, V, U^*, V^*)_i\}_{i=1}^{n \cdot n^*}$
Perform abduction with $\mathcal{M}$ and $\{V_i\}_{i=1}^n$ to get $\{U_i\}_{i=1}^n$.
for $i = 1, \ldots, n^*$ do
    Perform cross-world abduction with $P(U^* \mid U)$ to get one draw of $\{U_i^*\}_{i=1}^n$.
    Perform prediction with $\mathcal{M}$ and $\{U_i^*\}_{i=1}^n$ to compute $\{V_i^*\}_{i=1}^n$.
    Append $n$ predicted samples $\{(U, V, U^*, V^*)_i\}_{i=1}^n$ from $P(U, V, U^*, V^*)$ to collected data.
end for
return samples $\{(U, V, U^*, V^*)_i\}_{i=1}^{n \cdot n^*}$
```

Scenario 1 (Balanced Qualifications)**.** *In this scenario, qualifications X_1 and X_2 are both balanced across age groups. In other words, the recruited applicant pool has good representation of relevant qualifications across all age groups. Scenario 1 is generated as follows: $A \sim Bern(0.5)$, $Z_A \sim \mathcal{N}(A/2, 1)$, $Z_{A'} \sim \mathcal{N}(0, 1)$, $X_1 \sim \mathcal{N}(Z_{A'}, 1)$, $X_2 \sim \mathcal{N}(3Z_{A'}, 1)$, and $Y \sim \mathcal{N}(X_1 + X_2 + 2Z_A + Z_{A'} - 1, 1)$. The corresponding DAG is shown in Figure 3a.*[6]

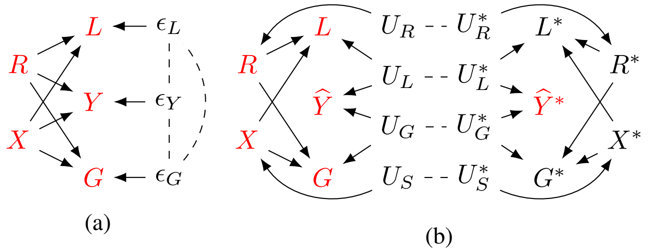
Scenario 2 (Unbalanced Qualifications)**.** *In this scenario, qualification score X_1 is not balanced across age groups. This imbalance could arise due to, for example, unequal access to developmental resources beforehand, i.e., age-related circumstance Z_A has impacted qualifications X_1. We generate data for Scenario 2 with a small modification to Scenario 1: $A \sim Bern(0.5)$, $Z_A \sim \mathcal{N}(A/2, 1)$, $Z_{A'} \sim \mathcal{N}(0, 1)$, $X_1 \sim \mathcal{N}(2Z_A + Z_{A'}, 1)$, $X_2 \sim \mathcal{N}(3Z_{A'}, 1)$, and $Y \sim \mathcal{N}(X_1 + X_2 + 2Z_A + Z_{A'} - 2, 1)$. The corresponding DAG is shown in Figure 3b.*

We fit several predictors on datasets of size 500 from the above equations, each a linear regression using some subset of the available covariates (excluding A). For each predictor, we obtain an estimate of $P(U, V, U^*, V^*)$ via Algorithm 1 using a non-informative backtracking conditional following Definition 1, and estimate the backtracking counterfactual terms of interest from $P(U, V, U^*, V^*)$, where we consider protected groups $A = 0$ and $A = 1$. We consider binarized outcome cases $\widehat{Y} > 0$ and $\widehat{Y} \leq 0$.

Figure 4 shows how well each of these models satisfies Definition 7 across the two scenarios. This is measured as

[5] Our code repository is located on GitHub at: https://github.com/lbynum/backtracking-counterfactual-fairness-and-recourse.

[6] In this simulation, the impact of A on Z_A is not viewed as a causal effect but instead as a convenient way of inducing dependence. This choice does not change any of the results, as no interventions on A are considered.



(a)

(b)

Figure 6: (a) A possible DAG for the law school example, reproduced from [Kusner *et al.*, 2017]. (b) A backtracking counterfactual twin network version of the DAG in (a), now with a Level 3 ICF fair predictor $\widehat{Y}$ as the outcome instead of Y.

the absolute energy distance maximum mean discrepancy (MMD) between the two terms — $P(S^* \mid \widehat{Y}^* = y^*, \widehat{Y} = y, X = x, A = a)$ and $P(S^* \mid \widehat{Y}^* = y^*, A = a')$ —*for each individual in the dataset* as a boxplot for each model. When these distances are zero for each individual, Definition 7 is satisfied. Figure 4 shows that in Scenario 1 with opportunity set $S = \{X_1, X_2, Z_{A'}\}$, a model using any subset of the mutable covariates (i.e., not using age A or age-related circumstance Z_A) will satisfy Definition 7. However, in Scenario 2, with the same opportunity set, no model satisfies Definition 7.

This comparison demonstrates an important takeaway. Even a random predictor will not balance an unbalanced predictor (like X_1) counterfactually across groups. The subsequent implication of even a random predictor being unfair is that, to achieve fairness, *we would have to change the underlying process that led to the imbalance in* X_1 *in the first place*. This connects to other discussions in, e.g., recourse literature on the need to consider *societal interventions* in addition to modifications of a classifier [Von Kügelgen *et al.*, 2022]. If, instead, imbalance in X_1 is deemed morally acceptable, and X_1 is thus not considered part of an individual's opportunity set S, then model performance follows a trend much like that for Scenario 1, where a model that doesn't make use of Z_A nor its now-descendant X_1 can easily satisfy Definition 7.

Figure 5 shows individual and group-level counterfactual effort required (Definitions 9 and 10) for the same datasets given opportunity set $S = \{X_1, X_2, Z_{A'}\}$ and MMD as cost function $\mathcal{L}$. Figure 5a shows how much effort is required of each individual in the dataset to get the opposite of their current outcome, showing wide variation in the cost of changing an outcome across different individuals, as well as more costly changes in Scenario 2 than in Scenario 1. Figure 5b shows how much effort is required instead for each group ($A = 0$ and $A = 1$) to flip their outcome starting from negative ($\widehat{Y} \leq 0$) or starting from positive ($\widehat{Y} > 0$), this time across 10 runs to account for variation in estimating MMD.[7] This figure demonstrates that even for the same model, group-level takeaways can often be the opposite of individual-level takeaways: at the group-level, Scenario 1 is often more costly than Scenario 2. Figure 5b also shows that changing outcomes is typically harder for group $A = 0$ than for group $A = 1$, across both scenarios.

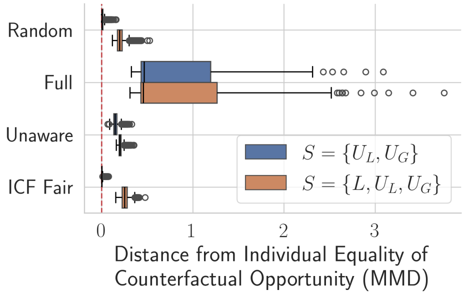


Figure 7: Visualization of Individual Equality of Counterfactual Opportunity on the law school dataset for four predictors: a random model (Random), a model using all available covariates (Full), an 'unaware' model using only L, G (Unaware), and a model satisfying interventional counterfactual fairness (ICF Fair). (a) Opportunity set $S = \{U_L, U_G\}$. (b) Opportunity set $S = \{L, U_L, U_G\}$.

These quantities — and our takeaways from using them — are immediately applicable in real data settings, like the one discussed in Section 5.2.

5.2 Law School Dataset

Used to illustrate counterfactual fairness in [Kusner *et al.*, 2017], the law school dataset from [Wightman, 1998] contains information on law students' demographics and academic performance compiled from a 1998 Law School Admission Council survey. Kusner *et al.* imagine a task of predicting the average grade of prospective students in their first year of law school in conjunction with the causal graph depicted in Figure 6a, where variables G, L, R, X, Y represent grade point average, entrance exam (LSAT) score, racial category, sex, and first-year average grade, respectively. Based on the structure of the model from [Kusner *et al.*, 2017], we have SCM $L = f_L(R, X) + \epsilon_L$; $G = f_G(R, X) + \epsilon_G$; $Y \sim \text{Bern}(p = \text{expit}(f_Y(R, X)))$; with linear functions f_L, f_G, f_Y and normally distributed ϵ_L, ϵ_G. For this dataset, focusing on R binarized to majority/minority and Y to high/low, Definition 7 compares $P(S^* \mid \widehat{Y}^* = y^*, R = r, X = x, G = g, L = \ell, \widehat{Y} = y)$ to $P(S^* \mid \widehat{Y}^* = y^*, R = r')$ for each individual.

One method to train a model that satisfies traditional, interventional counterfactual fairness (ICF) is to postulate a fully deterministic model with latent variables and use only the latent variables as inputs to the predictor. A 'Level 3' ICF fair predictor $\widehat{Y}$ fits separate regressions for L and G and uses residual estimates of ϵ_L, ϵ_G to predict Y. Figure 7 shows MMD distances (in the same manner as Figure 4) for this model and several others on a random sample of 5000 observations, with two different choices of opportunity sets. Distances are shown for the same sample using either $S = \{U_L, U_G\}$ or $S = \{L, U_L, U_G\}$. In the first case, both the ICF fair predictor and the random predictor satisfy Definition 7, highlighting that there are cases where existing bias mitigation techniques show promise in satisfying backtracking counterfactual criteria. In the latter case, however, we again see that none of the predictors satisfy Definition 7. This

[7] In this case, we have one sole computation per model rather than one for each individual in the dataset.

figure demonstrates that our choices about what imbalance is morally permissible or not (in this case, with LSAT scores) have relevance for finding a fair model not only in simulations but also in practice. These results also demonstrate empirically how the counterfactual discrimination criteria we introduce here not only capture something philosophically different from interventional criteria, but can also have different implications for the models we use.

6 Conclusion

In this work, we have introduced, to our knowledge, the first counterfactual-based fairness criteria that depart from the need to consider intervention on legally-protected characteristics, while still allowing us to consider fairness at an individual or a group level. We believe this paper can serve as a first step in enabling a new approach to applying counterfactual reasoning to demographic data for socially-relevant concepts like discrimination, charting a course for significant future exploration and technical development.

Acknowledgements

This research was supported in part by National Science Foundation (NSF) awards No. 1922658, 2312930, 2326193. We are grateful to Venetia Pliatsika and Falaah Arif Khan for helpful discussions.

References

[Balke and Pearl, 1994] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. *Probabilistic and Causal Inference*, 1994.

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

[Benthall and Haynes, 2019] Sebastian Benthall and Bruce D. Haynes. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 289–298, New York, NY, USA, 2019. Association for Computing Machinery.

[Bynum *et al.*, 2021] Lucius E.J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. Disaggregated interventions to reduce inequality. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021.

[Bynum *et al.*, 2023] Lucius E. J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. Counterfactuals for the future. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14144–14152, Jun. 2023.

[Carey and Wu, 2022] Alycia N. Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5, 2022.

[Chiappa, 2018] Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, 2018.

[Ehyaei *et al.*, 2023] Ahmad Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schölkopf, and Setareh Maghsudi. Robustness implies fairness in causal algorithmic recourse. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[Hanna *et al.*, 2020] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA, 2020. Association for Computing Machinery.

[Holland, 1986] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[Hu and Kohler-Hausmann, 2020] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 513, New York, NY, USA, 2020. Association for Computing Machinery.

[Jacobs and Wallach, 2021] Abigail Z. Jacobs and Hanna Wallach. *Measurement and Fairness*, page 375–385. Association for Computing Machinery, New York, NY, USA, 2021.

[Karimi *et al.*, 2022] Amirhossein Karimi, Gilles Barthe, Bernhard Scholkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55:1 – 29, 2022.

[Kasirzadeh and Smart, 2021] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 228–236, New York, NY, USA, 2021. Association for Computing Machinery.

[Kilbertus *et al.*, 2017] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[Kusner *et al.*, 2017] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, 2017.

[Loftus *et al.*, 2018] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *ArXiv*, abs/1805.05859, 2018.

[Nabi and Shpitser, 2017] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2018:1931–1940, 2017.

[Pearl, 2000] Judea Pearl. Causality: Models, reasoning and inference. 2000.

[Plecko and Bareinboim, 2022] Drago Plecko and Elias Bareinboim. Causal fairness analysis. *ArXiv*, abs/2207.11385, 2022.

[Sen and Wasow, 2016] M. Sen and Omar Wasow. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Review of Political Science*, 19:499–522, 2016.

[Von Kügelgen *et al.*, 2022] Julius Von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9584–9594, 2022.

[Von Kügelgen *et al.*, 2023] Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pages 177–196. PMLR, 2023.

[Wightman, 1998] Linda F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[Wu *et al.*, 2019] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Neural Information Processing Systems*, 2019.

[Zhang and Bareinboim, 2018] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. In *AAAI Conference on Artificial Intelligence*, 2018.