# Modeling Discrimination with Causal Abstraction

Milan Mossé* Kara Schechtman* Frederick Eberhardt Thomas Icard
*UC Berkeley* *Princeton* *Caltech* *Stanford*

**Abstract**

A person is directly discriminated against only if their possession of a protected attribute, such as gender or race, causes their worse treatment. This implies that a protected attribute is sufficiently separable from other attributes to isolate its causal role. But gender is embedded in a nexus of social factors that resist isolated treatment. If gender (or another protected attribute) is socially constructed, in what sense can it cause worse treatment? Some suggest that since causal models require *modularity*, i.e. the ability to isolate causal effects, attempts to causally model discrimination are misguided. Others suggest that we can preserve modularity by saying that the perception of gender, rather than gender itself, causes worse treatment.

We address this problem differently. We introduce a framework for reasoning about discrimination, in which gender is a high-level *abstraction* of lower-level features. In this framework, gender can be modeled as itself causing worse treatment. Modularity is ensured by allowing assumptions about how protected attributes are constituted to be precisely and explicitly stated, via an alignment between a protected attribute like gender and its constituents. Such assumptions can then be subjected to normative and empirical challenges, which lead to different views of when discrimination occurs. By distinguishing constitutive and causal relations, the abstraction framework pinpoints disagreements in the current literature on modeling discrimination, while preserving a precise causal account of discrimination.

## 1 Introduction

A person is directly discriminated against only if their possession of a protected attribute, such as gender or race, explains their worse treatment (Thomsen, 2018). When explanation is understood causally, we arrive at the notion of

> *Causal Discrimination:* A person is directly discriminated against only if their possession of a protected attribute, such as gender or race, causes their worse treatment.

The causal notion of discrimination is pervasive. It is invoked by U.S. discrimination law (Kohler-Hausmann, 2018), by leading algorithmic fairness criteria (Kilbertus et al., 2017; Kusner et al., 2017; Barocas et al., 2023; Plečko and Bareinboim, 2024; Loi et al., 2023), and by *audit studies*, in which experimenters detect institutional discrimination using techniques that resemble randomized controlled trials (Gaddis, 2018). For example, to test for gender discrimination in the job market, experimenters compare callback rates for resumes which are identical except for a gender-coded attribute, such as the applicant's name and pronouns, photo, or directly disclosed gender (Galos and Coppock, 2023).

However, even assuming we have settled on a theory of which attributes should be protected from discrimination and why it is wrongful (e.g. Hellman 2008; Kolodny 2023), the notion of causal discrimination is difficult to spell out precisely. Indeed, on a counterfactual understanding of causation (Lewis, 1973; Woodward and Hitchcock, 2003a,b; Woodward, 2005), to determine whether gender caused worse treatment, we must consider whether an individual's treatment would have been worse, had their gender and *only* their gender been different.[1] But one may wonder what exactly it means to intervene only on a person's gender, given that a person's gender seems bound up with their other attributes, such as their name and self-presentation.

After all, causal models require *modularity*, i.e. that no intervention on a variable constitutes an intervention on other model variables. But many *social constructionist* theories of gender posit that names and self-presentation are partially constitutive of gender (Haslanger, 2017; Hu and Kohler-Hausmann, 2020). For instance, names like 'Ann' form part of the social meaning of what it is to be gendered as a woman. As a result, an intervention setting a name of an applicant on a resume from "Ann" to "Andrew" cannot, in principle, be separated from an intervention on their gender, and vice versa. Placing gender and names in a causal model together—as formalizations of audit studies do, to justify using an intervention on names to reveal an effect of gender—requires making a false modularity assumption that names do not partially constitute gender. Thus a model with both names and gender satisfies modularity only if gender is an "intrinsic" or "essential" feature of an individual, which is not constituted by names or any other model variables (Hu and Kohler-Hausmann 2020, pp. 4, 9; Hu 2024b, p. 9).

Similarly, to hold constant the other features on a resume, such as college attended, auditors must place them in a causal diagram. But as we later explain, such attributes are plausibly partially constitutive of gender (for a motivating example, imagine the college attended is Barnard vs. Columbia). (Hu and Kohler-Hausmann 2020, pp. 4, 9; Hu 2024b, p. 9). A recent audit of large language models concludes by conceding that modularity raises a problem for modeling racial discrimination: "Conceptually, it is challenging to rigorously define what it would even mean to manipulate an LLM's perception of an individual applicant's race in isolation from other factors." (Gaebler et al., 2024)

More generally, Hu and Kohler-Hausmann (2020) raise a *modularity problem* for causal models of discrimination, which we propose to reconstruct as follows:

[1] We focus on structural causal models Pearl (2000), given that they are used in algorithmic fairness (Kilbertus et al., 2017; Kusner et al., 2017), and that their ability to accommodate racial discrimination has already been subject to criticism (Kohler-Hausmann, 2018; Hu and Kohler-Hausmann, 2020; Hu, 2024b) and defense (Weinberger, 2022). For discussion of analogous difficulties for the potential outcomes framework, see Glymour and Glymour (2014); Marcellesi (2013); Greiner and Rubin (2011); Sen and Wasow (2016).

1. *Causal discrimination:* A person is directly discriminated against only if they are treated worse than others, and this is caused by their protected attribute.

2. *Modularity:* Causal models require *modularity*, i.e. that no intervention on a variable constitutes an intervention on other model variables.[2]

3. *Constituents in the Model:* The constituents of protected attributes often need to appear in causal models of discrimination. For example, audit studies intervene on names, and social constructionism says that names constitute gender.

4. Therefore, causal models cannot express claims about discrimination: they would have to include protected attributes and their constituents, violating modularity.

This worry has prompted many to defend causal models, typically by arguing that causal models of discrimination do not require modularity for protected attributes such as gender or race, but rather for some distinct feature associated with those protected attributes. For example, Greiner and Rubin (2011) propose that racial discrimination occurs if others' *perception* of race causes worse treatment (cf. Singh and Wodak 2023), Sen and Wasow (2016) propose that race has several constituents and that racial discrimination occurs if any *constituent* causes worse treatment, and Weinberger (2022) proposes that gender discrimination occurs if *signals* of gender cause worse treatment.[3] We argue that the modularity problem is not adequately addressed by these proposals.

Those who developed the modularity problem conclude that we must think of protected attributes like race or gender as explanatory in some sense that cannot be represented by traditional causal models (Kohler-Hausmann, 2018; Hu and Kohler-Hausmann, 2020; Dembroff et al., 2020; Hu, 2024b). Like the alternatives of skepticism (Holland, 1986) or agnosticism (Tolbert, 2024a) that race or gender is a cause, this is in tension with the notion of causal discrimination deployed by audit studies, algorithmic fairness criteria, the U.S. law, and those describing their experience of discrimination (Cherry and Bendick Jr., 2018, pp. 45-46, 53-55).

Crucially, all of the above responses to the modularity problem share what we can call

[2] We thus define modularity as what Woodward (2015) calls "independent fixability" and Weslake (2024) calls "independent manipulability." There are many definitions of modularity in the literature. For example, *principle of independent mechanisms* says that for all variables $V$ in a causal model, the mechanism $f_V$ by which it depends on its causal parents and exogenous noise does not contain information about the distribution of any other variable $X$ (see Janzing and Schölkopf (2008); Peters et al. (2017) pp. 17-21, point 2 and pp. 58-62). This is sometimes called "modularity" but is distinct from the notion of modularity we discuss. We are simply adopting the terminology of those who raise the modularity problem (Hu and Kohler-Hausmann, 2020).

[3] In an influential audit study, Bertrand and Mullainathan (2004) describe themselves as manipulating the interviewer's perception of race, rather than race itself (Bertrand and Mullainathan, 2004, p. 992). However, audit studies have also been seen as establishing direct, causal claims about protected attributes (Cherry and Bendick Jr., 2018). The issue of whether these studies manipulate protected attributes or the perception of them is orthogonal to the one we take up here, since the modularity problem arises regardless(Hu and Kohler-Hausmann, 2024). Although our framework is technically agnostic between protected attributes and the perception of them, we focus here on the attributes. For discussion see §3.3.

> *the Single-Level Assumption:* When causally modeling some phenomenon, one must place all of the relevant variables in a single causal model, which describes the situation at a single level of abstraction.

We jettison the Single-Level Assumption, by modeling gender as a high-level abstraction of lower-level attributes. We introduce

> *the Abstraction Model for Discrimination:* A person's gender is constituted by "low-level" natural and social attributes and relations. Gender causes an outcome if a change in gender, instantiated by a change in the low-level attributes constitutively aligned with gender, causes a change in outcome. This is the sense in which gender causes worse treatment, when gender discrimination occurs.

This model allows assumptions about how protected attributes are constituted to be precisely and explicitly stated, via an alignment between the attribute and its constituents. By distinguishing constitutive and causal relations, the abstraction framework dissolves the modularity problem: in our model, the relationship between gender and its constituents remains non-modular, in the sense that one cannot intervene on gender without intervening on its constituents. But as we explain, this is a benign (indeed intended) feature of abstraction, not an obstacle to causal modeling: when one attribute is simply a higher-level abstraction of other, lower-level attributes, one should expect changes at one level to correspond to changes at the other. While we develop the model by applying it to gender discrimination in audit studies, it applies equally to other protected attributes, such as race, disability, and religion, and in other settings, including algorithmic fairness and the U.S. law.

After introducing the abstraction model of discrimination (Section 2) and addressing the modularity problem (Section 3), we use discuss how our model bears on audit studies (Section 4). We introduce an initial model of audit studies (Section 4.1) and distinguish several challenges for a successful model of audit studies (Section 4.2). In particular, we consider the worry that interventions which purport to test for discrimination in fact reveal quantities which are in some sense irrelevant to discrimination (Hu and Kohler-Hausmann, 2020). Stated precisely, this worry provides reasons for objecting to particular causal models of discrimination. While an account of which interventions are normatively relevant to discrimination lies beyond the scope of this paper, we argue that the need for such an account is not a problem for causal models of discrimination as such, because the relevant interventions and their shortcomings can be stated precisely using causal abstraction. In closing, we discuss the implications of our model for experimental design (Section 4.3).

# 2 Abstraction and discrimination

The social construction of protected attributes like gender implies that an individual's gender is not a feature on par with other natural or social attributes and relations, but rather defined in terms of such "lower-level" attributes and relations. While such multi-level representation is not part of a standard causal model, causal abstractions can represent causal systems at multiple levels of granularity. This section will introduce the framework of causal abstraction and then show how it can be used to model discrimination.

## 2.1 Introducing abstraction

Consider an experiment in which a bird is trained to peck at objects of any shade of red (Yablo, 1992; Woodward, 2021). When the bird pecks at some red object, say a crimson one, we can make two causal claims about it, at different levels of abstraction: that it pecks at the object because it is *crimson*, or that it pecks at an object because it is *red*. Each of these causal claims suggests an associated causal model, one in which perception of a fine-grained color (e.g., crimson, scarlet, cyan, turquoise) causes pecking behavior, and another in which perception of a coarse-grained color (e.g., red, blue) causes pecking behavior.

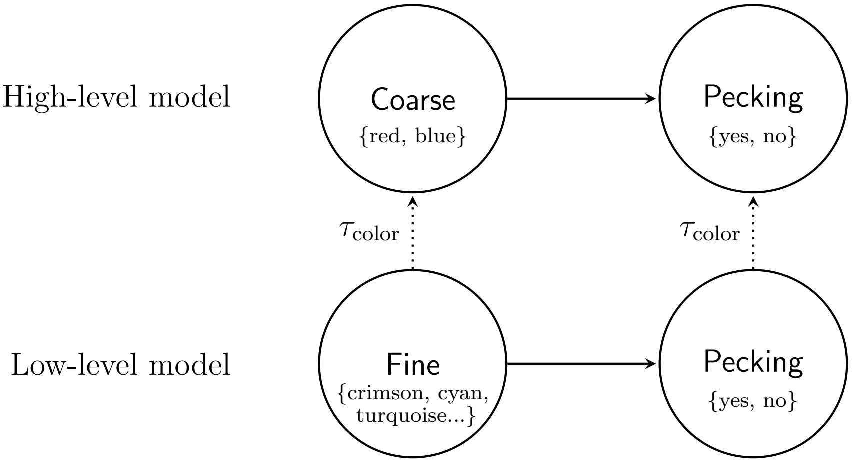


Figure 1: Color as an abstraction. Dotted arrows represent an alignment $\tau_{\text{color}}$ between causal models, which maps values of the Fine color variable (crimson, scarlet) to values of the Coarse color variable (red). Solid arrows express causal relationships, e.g. that intervening on the value of the Coarse color variable can change the value of the Pecking variable.

The fine-grained and coarse-grained models are depicted in Figure 1. The high-level model contains two variables: Coarse, which represents the bird's perception of a coarse-grained color, and Pecking, which represents whether the bird pecks. The causal arrow implies that Pecking is an effect of Coarse, and thus is some function of Coarse, perhaps together with random noise (included to capture variation among birds' pecking behavior not explained by object color). The low-level model is almost exactly the same, except it models pecking behavior as a function of fine-grained color Fine rather than coarse color.

Let us introduce the relevant notion of a causal model more formally.

**Definition 1.** A *structural causal model* is a tuple $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P} \rangle$, where $\mathbf{V}$ are the model variables, $\mathbf{U}$ are the noise variables, $\mathbb{P}$ is a distribution over $\mathbf{U}$, and $\mathcal{F}$ are the structural equations. The value of each noise variable $U \in \mathbf{U}$ belongs to $\text{Val}(U)$ and is determined by $\mathbb{P}(\mathbf{U})$. The value of each model variable $X \in \mathbf{V}$ belongs to $\text{Val}(X)$ and is determined by the structural equation $f_X \in \mathcal{F}$ which is a function of the values of other variables:

$$X := f_X(\mathbf{V} \setminus \{X\}, \mathbf{U}).$$

Evaluating the structural equations according to the distribution of noise given by $\mathbb{P}$ extends it to an *observational distribution* over $\mathbf{V}$.

In other words, a causal model contains variables $X_1, X_2, X_3 \dots,$ where the value of each variable $X$ is determined as a function $f_X$ of the values of other variables, and perhaps some random noise, which captures variation not explained by model variables (see, e.g., Pearl 2000, Section 1.4.1; Peters et al. 2017, Definition 6.2; Bareinboim et al. 2022, Definition 1).

The effect of a manipulation on a variable is modeled as an *intervention* $\text{do}(X = x)$, which replaces the function $f_X$ defining $X$ with a fixed value $x$ (Pearl 2000, Section 1.4.3; Peters et al. 2017, Definition 6.8; Bareinboim et al. 2022, Definition 5).Because other variables may be defined as functions of $X$, the intervention $\text{do}(X = x)$ percolates through the model, changing the likelihood that other variables take various values; this is the effect of the intervention. This gives us an *interventional* distribution on other model variables.

**Definition 2.** Given a causal model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P} \rangle$, an *intervention* on a model variable $I = do(X_i = x_i)$ creates a new causal model $\mathcal{M}^I = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}^I, \mathbb{P} \rangle$, which is identical to $\mathcal{M}$ except its structural equations, with $f^I_{X_i} = x_i$ and $f^I_{X_j} = f_{X_j}$ for $j \neq i$. The *interventional distribution* $\mathbb{P}^I$ denotes the joint distribution over $\mathbf{V}$ by extending $\mathbb{P}$ over $\mathbf{V}$ with $\mathcal{F}^I$. The definition can be applied to interventions on multiple variables (which replace multiple structural equations in $\mathcal{F}$ with constants).

We are now in a position to explain the sense in which the high-level model "simplifies" the low-level one. Consider two models $\mathcal{M}_{\text{low}}$ and $\mathcal{M}_{\text{high}}$. An *alignment* $\tau$ is a function which maps values of variables in the low-level model $\mathcal{M}_{\text{low}}$ to values of variables in the high-level model $\mathcal{M}_{\text{high}}$ (Geiger et al. 2025, Definition 31; Rubenstein et al. 2017, Section 4.2; Beckers and Halpern 2019, $\tau$ in Definition 3.1 ff.):

**Definition 3.** An *alignment* $\tau$ between the causal models $\mathcal{M}_{\text{low}}$ and $\mathcal{M}_{\text{high}}$ maps values of variables $X_1, \dots, X_k$ in $\mathcal{M}_{\text{low}}$ to values of variables $Y_1, \dots, Y_\ell$ in $\mathcal{M}_{\text{high}}$:

$$\tau(X_1 = x_1, \dots, X_k = x_k) = (Y_1 = y_1, \dots Y_\ell = y_\ell).$$

Conceptually, an alignment may be thought of as a *simplification* of the low-level model: for example, it may be constructed by mapping different values of a low-level variable to individual values of that variable (Geiger et al., 2025, Definition 38), mapping values over multiple low-level variables into values of a single variable (Geiger et al., 2025, Definition 37), or ignoring the values of some low-level variables entirely (Geiger et al., 2025, Definition 36), or a combination of these operations. In the above example, the alignment $\tau_{\text{color}}$ merges the fine color values into the coarse color values by mapping crimson and scarlet to red, while mapping cyan and turquoise to blue. As in this example, the alignment $\tau$ need not be injective, and as a result, interventions on high-level variables will often be ambiguous between several different interventions on low-level ones. (We discuss this issue further in Section 4).

So-defined, alignments are highly permissive: strictly speaking, any model can be aligned with any other. As we now explain, to define an *abstraction*, an alignment must also ensure that interventions in the low-level model and interventions in the high-level model are *causally consistent*. To check whether $\tau_{\text{color}}$ is causally consistent, we first perform an intervention on color at the low level, and then transform the effect it has on pecking to the high level via the alignment. We then check that this has the same result as first transforming the intervention via the alignment, and then intervening at the high level. This condition is in fact met by the alignment $\tau_{\text{color}}$, because intervening on fine-grained color by changing it from crimson to cyan and then checking whether the bird pecks has the same result as intervening on coarse-grained color by changing it from red to blue and then checking whether the bird pecks. By mapping more general colors to corresponding more specific shades, $\tau_{\text{color}}$ tracks a real metaphysical relationship between a color and its many shades, to which pecking behavior is insensitive. More generally, an alignment $\tau$ between two models is causally consistent if intervening on variables in the low-level model $\mathcal{M}_{\text{low}}$ and then aligning into the high-level model $\mathcal{M}_{\text{high}}$ is the same as first aligning into the high-level model and intervening there (Rubenstein et al. 2017, Definition 3; Geiger et al. 2025, Definition 25):

**Definition 4.** An alignment $\tau$ between causal models $\mathcal{M}_{\text{low}}$ and $\mathcal{M}_{\text{high}}$ is *causally consistent* if intervening at the low level and then aligning to the high level is the same as first aligning the low-level intervention and then performing it at the high level. In other words, where $\mathbb{P}^{I}_{\text{low}}$ denotes the interventional distribution over the values of all low-level variables that result from intervention $I$, and $\mathbb{P}^{\tau(I)}_{\text{high}}$ denotes the interventional distribution over high-level values that results from intervention $\tau(I)$:

$$\tau(\mathbb{P}^{I}_{\text{low}}) = \mathbb{P}^{\tau(I)}_{\text{high}}.$$

Given such an alignment, we say that $\mathcal{M}_{\text{high}}$ is an *abstraction* of $\mathcal{M}_{\text{low}}$.

Causal consistency is a stringent requirement, requiring perfect correspondence between low-level and high-level interventions. We introduce it for simplicity, but it can be relaxed in various ways (see, e.g., Beckers et al. 2020; Rischel and Weichwald 2021; Geiger et al. 2025, Definition 41). Most simply, we can say that $\tau$ is *$\epsilon$-causally consistent* when for some $\epsilon$ and all interventions $I$,

$$D(\tau(\mathbb{P}^{I}_{\text{low}}),\, \mathbb{P}^{\tau(I)}_{\text{high}}) < \epsilon,$$

where $D(\mathbb{P}_1, \mathbb{P}_2)$ is a distance measure between probability distributions, e.g. total variation (cf. Plečko and Bareinboim 2024, Appendix D1).

## 2.2 Modeling gender with an alignment

We now explain how to use an alignment to model the constitutive relations between protected attributes, such as gender, and their constituents. Any theory of what constitutes gender can be "plugged in" to the framework we present. However, we focus here on the view that gender is *constitutively socially constructed* (Mallon, 2024), in the sense that gender status is largely constituted by social facts, such as facts about social relations or social status (Haslanger, 2017, p. 20).

The view that gender is socially constructed is attractive and plausible because it is thought to make good sense of transgender and genderqueer identities (Jenkins 2016; Dembroff et al. 2020, p. 4), as well as socially-determined variation in how intersex individuals are gendered (Ásta, 2013, pp. 726-7). However, our goal in this paper is not to provide a defense of social constructionism about gender. Instead, we are simply granting the view that gender is socially constructed to those who raise the modularity problem (Hu and Kohler-Hausmann, 2020) and then arguing that our framework can accommodate this view and dissolve the modularity problem.

In particular, we grant that attributes like the following, which have been widely discussed in the literature on the modularity problem, are constitutive of gender statuses:

- Attributes of an individual's forms of address, such as their *name* or *pronouns* (Kohler-Hausmann and Dembroff, 2022, p. 90).
- Attributes of an individual's self-presentation, such as *wearing a skirt* or *makeup* (Hu, 2024a, p. 8).
- Attributes of an individual's education, such as *whether they attended a women's college* (Hu, 2024a, cf. p. 11).

The idea is not that any change in one of these attributes is necessary or sufficient for a change in gender, but rather that some such changes, often together with others, constitute a change in gender. For example, if a person who initially identified as a man were to change their name from "Andrew" to "Anne," adopt she/her pronouns, begin wearing a skirt and makeup, and enroll in an all-women's university, they would plausibly have transitioned.

The relations between gender and its constituents induce an alignment $\tau_{\text{gen}}$ that specifies how attributes constitute gender (Figure 2). To be sure, this already involves a substantive modeling assumption. One might hold that it is impossible, even in principle, to specify how exactly gender is constructed from lower-level attributes of individuals. For example, one might claim that gender, socioeconomic class, and race are all mutually constituted by each other, rather than by a more fine-grained basis of lower-level attributes.[4] However, we will set this possible worry aside, because (1) this claim is potentially consistent with the abstraction model, insofar as we may model gender as abstracting race in one context, and model race as abstracting gender in another; and (2) we assume that social constructionist concerns regarding the possibility of causal models of discrimination do not rest on thoroughgoing skepticism that a social construction of gender in terms of lower-level attributes could ever be articulated in principle.

The simple assumption that gender aligns with these lower-level attributes already licenses answers to counterfactual questions about gender. For instance, it makes sense to ask what would happen were one to intervene on a person's gender in a specific way. This is just to ask what would happen, were one to intervene on the lower-level attributes of an individual in such a way that their gender would now be different, according to the constitution of gender, as defined by the alignment $\tau_{\text{gen}}$. (We discuss whether this alignment is causally consistent, and thus defines an abstraction, in §4.)

[4] For arguments that would make room for this possibility, e.g., by rejecting the traditional view that constitutive relations are asymmetric and irreflexive, see Barnes (2018) and Jenkins (2016), respectively. See also Bright et al. (2016) and Wang et al. (2022) for further discussion about formally modeling intersectionality.

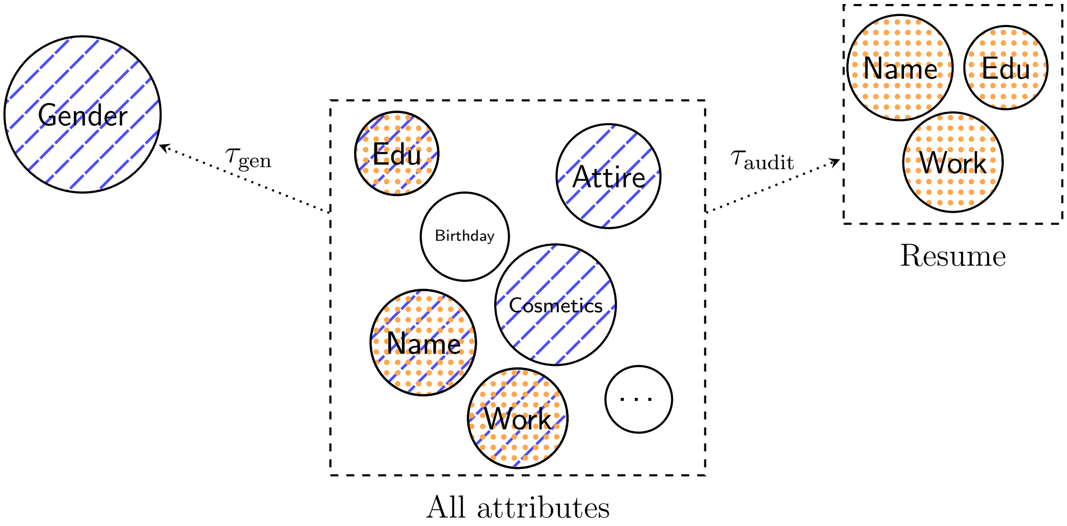


Figure 2: Gender and resume as abstractions. The alignment $\tau_{\text{gen}}$ comes from a theory of how gender is constituted, and the alignment $\tau_{\text{audit}}$ is used by audit studies. The dotted box labeled "all attributes" contains all attributes of an individual, some of which are abstracted into gender and marked with blue lines going northeast (e.g., their attire, cosmetics, name, work, education in the diagram), and others of which are discarded by the alignment $\tau_{\text{gen}}$ (e.g. the day of the year somebody is born in the diagram). The dotted "Resume" box represents a collection of variables from all attributes, including work history, education, name, and so on; these are marked with orange dots. The ellipses above indicate that many attributes which might belong on a resume or in the constitutive basis of gender are not pictured.

# 3 The modularity problem

In this section, we explain the modularity requirement and the modularity problem for social constructionism in causal models of discrimination. We then explain how the abstraction framework for modeling discrimination addresses this problem and compare it to some alternatives.

## 3.1 Introducing the modularity problem

To ensure that interventions on individual variables reveal the effects of those variables, structural causal models must satisfy a modularity requirement (see, e.g., Pearl 2000, p. 63; Woodward 2005,p. 327; Peters et al. (2017), pp. 17-20, point 1). Modularity requires that no intervention on any set of variables constitutes an intervention on variables outside that set. For example, a causal model which implies that object color causes bird pecking behavior is modular because interventions on each of the two variables do not constitute interventions on the other. If we paint an object blue, we perform an intervention that directly changes just that object's color, not anything about the bird.[5] As a result, any effects of this intervention

[5] Of course, since variables often causally influence other variables, modularity allows for an intervention on some variable to affect other variables through *causal* relations. Interventions on the color of the object plainly affect the bird's actual pecking causally, but color and pecking behavior are still modular. Instead,

demonstrate the causal relationship between object color and bird pecking behavior.

Without modularity, the causal effects of an intervention on a variable are hopelessly conflated with the intervention itself. Consider, for instance, a causal model claiming that the coarse color of an object causes its particular fine shade. This model is non-modular because it is conceptually impossible to manipulate the color of an object separately from intervening on its particular shade (and often, vice versa). How could we change the color of an object from red to blue, without changing it from some shade of red to some shade of blue? Since interventions on fine shades constitute interventions on coarse colors, we cannot isolate a causal effect of the intervention.

The modularity problem, developed by Lily Hu and Issa Kohler-Hausmann (Kohler-Hausmann, 2018; Hu and Kohler-Hausmann, 2020), is that attributes like race, gender, and religion cannot be placed in a single causal model with the attributes from which they are socially constructed, because their constitutive relations violate the modularity requirement. In the case of audit studies, the problem is that gender and names cannot be placed in a single structural causal model without violating the modularity assumption crucial to causal modeling and inference. The worry is that these two variables interact more like colors and their more specific shades than like bird pecking behavior and object colors; models including both variables inevitably end up non-modular.

The problem is easiest to see if we accept that gender is a social construction, and thus stands in *constitutive* rather than *causal* relations to many of its correlates. For then certain names constitute part of the social meaning of gender, and thus interventions switching names from "Anne" to "Andrew" constitute interventions to gender itself. This problem arises for any attributes that appear on a resume, like education or employment history, whenever they partly constitute gender, yet causal tests of discrimination often purport to hold these attributes fixed while manipulating gender.

## 3.2 Abstraction as a solution to the modularity problem

The abstraction model of discrimination addresses the modularity problem. Because socially constructed attributes and their constituents are placed in *separate causal models*, and modularity need only hold within each model taken separately, there is no non-modularity within either the high-level model containing gender or the low-level model containing its constituents.

It is sometimes objected that structural causal models of discrimination cannot accommodate constitutive relations (Hu and Kohler-Hausmann, 2020, p. 5). This objection can be overcome with abstraction: the alignment nicely captures the relevant constitutive relations.

By dissolving the modularity problem, the abstraction model immediately dissipates the concern that it is impossible to causally model gender in settings where its constituents are also endogenous (e.g., discrimination auditing). This in turn shows that the modularity requirement does not in itself force causal models to posit an especially implausible metaphysical boundary between gender and the context in which it is socially embedded, e.g.

when a variable $X$ is modular with respect to $Y$, this means that interventions on $X$ do not constitute interventions on $Y$. While this distinction between causal and non-causal relations is an intuitive one, Janzing and Mejia (2022) discuss a number of puzzles about how to make it fully precise.

by treating gender as an “intrinsic” or “essential” feature of individuals (Hu and Kohler-Hausmann 2020, pp. 4, 9, Hu 2024b, p. 9). Instead, the relationship between gender and other factors is modeled within the framework, via the alignment $\tau_{\text{gen}}$ that specifies how gender is socially constructed. This alignment may include all the attributes that are included in a particular social constructionist theory of gender. Although it excludes all constituents of gender, the high level model containing gender may also include variables representing causes and effects of gender.

We emphasize that the abstraction framework does not entail that gender, race, or other protected attributes are socially constructed. Instead, it provides a way of modeling the claim that these attributes are socially constructed, and in doing so addresses a modularity worry that arises for social constructions. Though we find it plausible that at least some protected attributes are socially constructed, our argumentative strategy is to simply grant to those who raise the modularity problem that protected attributes are socially constructed, and then to argue that this is not an obstacle to causally modeling them.

## 3.3 Comparison to other approaches

Our proposal develops Sen and Wasow’s (2016) suggestion to treat social constructions (in their case, race) as a “bundle of sticks,” such that manipulations on race are manipulations on some of the sticks. Our idea is to spell out this suggestion using an alignment $\tau_{\text{gen}}$ between two causal models, one with the constituents of gender and another with gender itself. We show in the next section how the framework allows us to distinguish several desiderata for a successful model of audit studies.

Concerns about protected attribute’s manipulability lead Greiner and Rubin (2011) to propose that we use the perception of race, rather than race itself, as a treatment in causal studies of race. Some audit studies even explicitly frame themselves as intervening on perception (Bertrand and Mullainathan, 2004). But replacing protected attributes by their perception simply pushes the modularity problem one step back: if race or gender is socially constructed, we should expect interventions on their perception to be bound up with interventions on the perception of names, education, and so on (cf. Hu and Kohler-Hausmann (2024)).[6] Further, as we saw, U.S. discrimination law explicitly invokes the idea that attributes like sex or race themselves cause worse treatment—so there is value in seeing if this idea can be made to work, if want models of discrimination that can invoked in the courtroom.

Weinberger (2022) suggests, more generally, that we can model audit studies as manipulating a signal of gender that is causally downstream of gender. Audit studies vary names “as if” they were varying gender, thereby revealing an effect of gender along a particular causal path.[7] To include gender and names in the same causal model without violating modularity,

[6] As Weinberger (2022) observes, it is unclear that when a victim is harmed based on a mistaken perception of their protected attribute, we should count this as wrongful discrimination. Whether a person actually possesses the attribute in question may inform the way in which the act is wrong, and the degree to which it is wrong. This is another worry for replacing the attribute with its perception.

[7] VanderWeele and Robinson (2014) propose a model of race as a cause which is very similar to Weinberger’s, in which socioeconomic status, physical phenotype, parental physical phenotype, genetic background, cultural context, family, and neighborhood function as signals of race (in Weinberger’s sense). This

Weinberger argues that it does not follow from the assumption that race is socially constructed that gender is constituted by its signals (e.g., names). If gender is not constituted by its signals, then it is possible to manipulate those signals separately from gender.

If this is going to work, the appeal to signals will need to model *every* relevant would-be source of modularity violations included in the model as a signal (rather than a constituent) of gender. Our model will need to include the signals we intervene on, as well as those we hold fixed. So not only names and pronouns, but also self-presentation, university, field of work, relations of subordination, and so on will need to be called signals (not constituents) of gender. When this is done, we seem to end up nothing at all left in the constitutive basis. We worry that this just rejects the view of social constructionism that motivates the problem we wanted to solve, which proposes that there are constitutive relations between protected attributes and many of their correlates (Hu, 2024, §4; Hu and Kohler-Hausmann, 2024, pp. 249-51). By contrast, the abstraction model does not require us to reject the social constructionist views that motivate the modularity problem: it allows that gender may be partly constituted by signals like names.[8]

# 4 Modeling audit studies

The abstraction model addresses the modularity problem, but the use of abstraction does not obviate the need to make normative judgments about the relevant counterfactual contrasts in testing the effects of social kinds. Instead, the abstraction model provides a framework for making these judgments precise and explicit, primarily via an alignment between higher-level protected attributes and their lower-level constituents. In this way, the abstraction model offers the flexibility to encode different assumptions about how discrimination should be measured, which lead to different verdicts about whether discrimination has occurred. As we now show, once these assumptions are made precise and explicit, they can then be subjected to further normative and empirical challenges.

We introduce an initial model of audit studies (Section 4.1) and distinguish several challenges for a successful model of audit studies (Section 4.2). In particular, we consider the worry that interventions which purport to test for discrimination in fact reveal quantities which are in some sense irrelevant to discrimination (Hu and Kohler-Hausmann, 2020). Stated precisely, this worry provides reasons for objecting to particular causal models of discrimination. However, we argue that it is not a problem for causal models of discrimination as such, because it can be stated within the framework of causal abstraction. In closing, we discuss the implications of our model for experimental design (Section 4.3).

proposal faces the same worries, but we focus on Weinberger's view because it is framed as a response to the modularity problem.

[8] Weinberger proposes that in many narrow experimental contexts, signals like names send sufficiently "isolated" signals of individuals' race, such that modularity can be adopted as a limited modeling assumption (Weinberger, 2022, p. 1282). We worry that this stipulation is overly optimistic, given the empirical evidence (summarized by Hu 2024) that racial perception, even if induced just by a name, starts a cascade of racialized interpretation of other attributes that at first blush do not convey racial meaning, like what university applicants attended, their hometown, or their work experience.

## 4.1 Modeling audit studies with abstraction

We can use the abstraction framework to propose a causal model of audit studies, by introducing two high-level models and alignments. Suppose some underlying causal model describes how low-level features of an applicant influence interview callback decisions. Audit studies select a strict subset of an individual's features, namely those appearing on their resume, "screening off" the effects of all other attributes of the individual on the outcome. In the terminology of abstraction, this defines an alignment $\tau_{\text{audit}}$, which discards any attributes not appearing on the resume and trivially aligns all attributes appearing on the resume with themselves (Figure 2). The resume is assumed to include all attributes of an applicant that exert a causal effect on interview status.

We can use the alignments $\tau_{\text{gen}}$ and $\tau_{\text{audit}}$ to state precisely the quantity of interest in common audit studies. A typical audit study fixes two resumes, $\text{Resume}_1$ and $\text{Resume}_2$, which differ only in their names. Because resumes only contain a few pieces of information about an individual, many different people with different attributes could possess the same resume. Where a "person" refers to a maximally descriptive set of attributes, let $\mathcal{A}_1$ be the set of people with resume $\text{Resume}_1$, and let $\mathcal{A}_2$ be the set of people with resume $\text{Resume}_2$. (In other words, $\mathcal{A}_1$ is the set of people that $\tau_{\text{audit}}$ aligns with $\text{Resume}_1$, and $\mathcal{A}_2$ is the set of people that $\tau_{\mathsf{audit}}$ aligns with $\text{Resume}_2$.)

The gender compositions of $\mathcal{A}_1$ and in $\mathcal{A}_2$ will tend to differ, especially given that the names appearing on the resumes $\text{Resume}_1$ and $\text{Resume}_2$, like Andrew and Anne, are highly correlated with gender. Let $\text{Gen}_1$ indicate the gender most common in $\mathcal{A}_1$ and let $\text{Gen}_2$ indicate the gender most common in $\mathcal{A}_2$.[9] Then it follows from the assumption that $\tau_{\text{gen}}$ and $\tau_{\mathsf{audit}}$ are causally consistent that audit studies successfully test the causal effect of gender on interview outcomes, in the following sense: the difference in interview status that results from changing $\text{Gen}_1$ to $\text{Gen}_2$ is the same as the difference that results from changing $\mathcal{A}_1$ to $\mathcal{A}_2$, which is in turn the same as changing $\text{Resume}_1$ and $\text{Resume}_2$. The necessary assumptions of causal consistency are licensed by the social constructionist theory of gender that defined $\tau_{\text{gen}}$, and by construction of audit studies, which omit all attributes not on the resume via $\tau_{\text{audit}}$ (Figure 3).

Formally, audit studies purport to test the effect of gender on interview status via the following equation:

$$\begin{aligned} & \mathbb{P}[\textsf{Interview Callback} \,|\, \text{do}(\text{Gen}_1)] - \mathbb{P}[\textsf{Interview Callback} \,|\, \text{do}(\text{Gen}_2)] \\ = \; & \mathbb{P}[\textsf{Interview Callback} \,|\, \text{do}(\text{Resume}_1)] - \mathbb{P}[\textsf{Interview Callback} \,|\, \text{do}(\text{Resume}_2)]. \end{aligned}$$

Admittedly, this picture is an over-simplification. There are multiple interventions on all of an individual's attributes which could correspond to a change from $\text{Gen}_1$ to $\text{Gen}_2$, so there is significant under-determination and ambiguity in finding a lower-level intervention to test the effect of this change. But this ambiguity is a common feature of causal inference (Spirtes and Scheines, 2004): in saying that we changed a stimulus for the bird from red to blue, it is ambiguous whether we changed it from crimson to cyan, or from scarlet to

[9] We align resumes to the most common gender in the population for simplicity, but other ways of aligning resumes to genders are certainly possible, such as probabilistically aligning $\text{Resume}_1$ to $\text{Gen}_1$ and $\text{Gen}_2$ according to the frequencies of each in the population $\mathcal{A}_1$ (and handling $\text{Resume}_2$ alignments analogously).

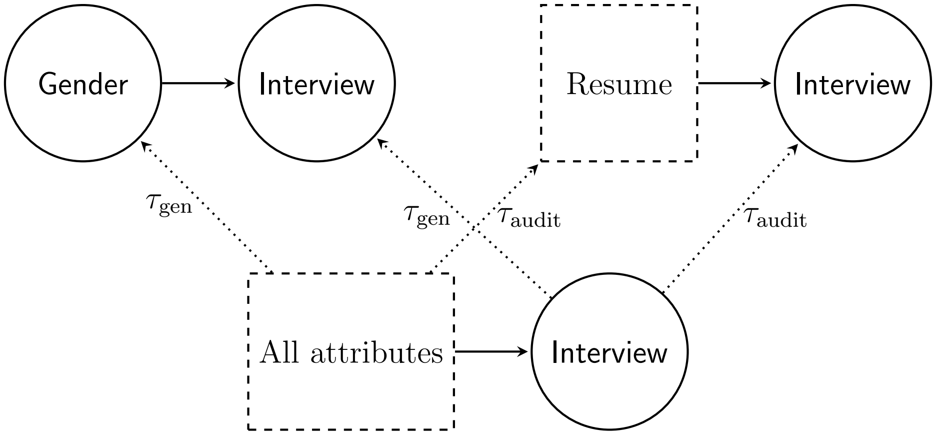


Figure 3: The abstraction framework applied to audit studies. The low level is represented by one causal model containing all modular, low-level attributes that affect interview callback decisions. The high level has two separate causal models, requiring two different alignments from the low level model. The first alignment, $\tau_{\text{audit}}$, is constructed by mapping all low-level attributes to those attributes that appear on the resume. Audit studies assume the high-level model, $\mathcal{M}_{\text{audit}}$, which contains only features on the resume, losslessly captures the causal information in the low-level model, since the hiring manager only sees the applicant's resume; this is sufficient for $\tau_{\text{audit}}$ to be causally consistent. The second alignment, $\tau_{\text{gen}}$, represents how gender is constituted by low-level attributes, aligning to a high level model $\mathcal{M}_{\text{gender}}$ capturing gender's effect on hiring. In addition to drawing from features on the resume, $\tau_{\text{gen}}$ may draw on attributes left out by $\tau_{\text{audit}}$ (Figure 2). However, under the assumptions that $\tau_{\text{gen}}$ is causally consistent and $\tau_{\text{audit}}$ is lossless, any intervention that changes gender by manipulating only attributes on resumes—such as toggling the applicant's name—will perfectly capture the effect of gender on interview status. (It is an open question whether this highly restricted intervention on Gender indeed is causally consistent.) Finally, audits construct two resumes that differ in gender-coded names and then measure their effect on outcomes. The goal is to measure the effect of gender through an intervention on resumes. Modeling the intervention as on resumes rather than gender allows experimenters to hold fixed the attributes on the resume that are *not* abstracted into gender (for example, an applicant's telephone number).

turquoise. Causal consistency guarantees that all such changes produce the same pattern of changes in outcome. Thus, to the extent that we have causal consistency of $\tau_{\text{gen}}$ in a diagram like Figure 3, there is a clear and principled sense in which audit studies can be modeled as testing the effect of interventions on gender. More generally, one can model claims about causal discrimination as claims about the effects of gender, understood as an abstraction of lower-level, in-principle manipulable attributes.

The abstraction framework thus addresses the objection that audit studies cannot, in principle, be modeled causally because of constitutive relationships between protected attributes and the attributes constituting them (Kohler-Hausmann, 2018). However, in the next section we use the framework of causal abstraction to distinguish some further challenges for modeling audit studies.

## 4.2 Challenges of intervention identification

We can now use the framework of causal abstraction to distinguish two challenges to the claim that by intervening on names, audit studies reveal a causal effect of a protected attribute that constitutes discrimination, which we discuss in turn:

- *Causal consistency (§4.2.1).* There must exist an alignment between gender and its lower-level constituents, such that testing effects of gender-coded names provides information about the overall effect of gender. At the extreme, this could be achieved if the alignment were perfectly causally consistent, since in that case, each intervention in the low-level model which corresponds (via the alignment) to a change in gender would have the same effect.

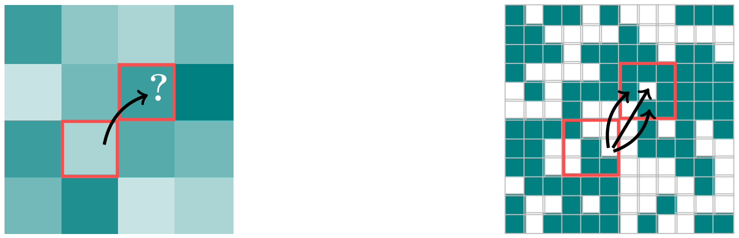
- *Normative relevance (§4.2.2).* The interventional quantities revealed by audit studies must correspond to gender subpopulations which are normatively relevant. For instance, a gender subpopulation might be relevant because it is in some sense "typical," or because it has other features of normative relevance to expectations of similar treatment. (To reveal the *absence* of a causal effect of gender, the interventional quantities revealed by audit studies would have to correspond to *all* the gender subpopulations that are normatively relevant.[10])

### 4.2.1 Causal consistency

Consider, first, a set of challenges associated with defining the alignment and high-level model. Interventions on a protected attribute will be ambiguous between interventions on its many constituents (cf. Tolbert 2024a, pp. 1103-5), which may have different effects. Indeed, even if the changes to the names on resumes performed in audit studies correspond, via the alignment $\tau_{\text{gen}}$, to changes in gender, there are many such changes. For example, changing university from Barnard to Columbia would typically also change the gender most likely associated with a given resume, and may well have a different effect on interview status than changing the name from "Anne" to "Andrew" (Figure 4).

This ambiguity does not arise if the alignment $\tau_{\text{gen}}$ is perfectly causally consistent, for then the effects of gender correspond perfectly with those on its aligned constituents. However, once we relax the stringent requirement of causal consistency, and allow that the effect of gender on interview status is some complex function of the varying effects of many different

[10] Many additional assumptions, often unwarranted, but unrelated to intervention identification, are needed to move from an audit's failure to confirm a causal effect of race or gender to a conclusion that there is not discrimination, because such a finding does not mean race or gender are not a cause of outcomes in some salient way. Narayanan (2022) outlines many obstacles to detecting these causal effects quantitatively, which can be illustrated by the resume audit study. For instance, a resume audit might be underpowered to detect very small differences in interview callback outcomes; yet these undetectable small differences can accumulate to large group disadvantages in the long run, leading to discrimination invisible in the "snapshot" dataset of an audit (Narayanan, 2022, pp. 9-10). Moreover, if most job discrimination occurs at the interview stage, then such discrimination will be invisible to the resume audits, which only looks at effects of race on callback rates (Narayanan, 2022, pp. 13-15). While posing fundamental challenges to the use of audits to substantiate claims of the absence of discrimination, these challenges do not originate from puzzles about taking race or gender as a cause in principle, but rather from data availability and hypothesis formation about the points in a decision making process at which race or gender can act as a cause.



(a) Coarse-grained

(b) Fine-grained

Figure 4: Measuring how an outcome could be different when moving from one category to another may depend on how precisely one picks exemplars of the two categories. Given a two-dimensional feature space (given by $x$ and $y$ coordinates), suppose settings of those features are clustered into categories, as boxes in Subfigure (a). Shade of teal denotes likelihood of an outcome (say, interview status) for a given category. Subfigure (b) reveals a finer-grained picture of the space of individuals, at which level outcome status is binary (yes or no). Taking an individual from one category to another can be done in any number of different ways, and these different possibilities may also differ with respect to the outcome. To the extent that the alignment is causally consistent, this type of ambiguity will be absent.

lower-level changes, we face the difficult tasks of (a) aggregating those low-level contrasts into a single, higher-level one and (b) measuring causal consistency. (In the case of audit studies, this involves defining the equation that underlies the arrow $\mathsf{Gender} \rightarrow \mathsf{Interview}$ in Figure 3, and thus deciding to what extent differential treatment on the basis of gender-coded names fully captures discrimination on the basis of gender.)

This aggregation function and causal consistency measure, like the alignment $\tau_{\text{gen}}$, will need to be informed by our understanding of the normative significance of discrimination. For example, obvious aggregation techniques, such as taking means, reflect a normative assumption that theories ought to capture the average experience. One of intersectional theory's founding texts—Kimberlé Crenshaw's critique of causal definitions of discrimination used in the law—questions this focus on averages, suggesting that we ought instead to look to the experiences of especially marginalized women, namely Black women, to understand discrimination based on gender (Crenshaw, 1989). Consequently, aggregation functions and causal consistency measures informed by intersectional critiques may not use simple averages. For example, one might say that the effect of a high-level intervention that changes gender from man to woman is the *geometric mean* (or the *maximum*) of the effects of all low-level interventions that realize this change in gender.

Similarly, whether an alignment is causally consistent "enough" is both a normative and empirical question. To obtain a theory of gender that is explanatory across a wider variety of contexts, or one that employs fewer high-level categories, we might have to relax our definition of approximate causal consistency further, by changing $D$ and $\epsilon$ (cf. §2.1)—which may be justified by different standards of consistency for sweeping social explanations than local explanations. Empirically, this hangs on whether gender proves to be an explanatory way of clustering lower-level attributes, once an aggregation function and measure of causal consistency have been chosen.[11] Current work in social science employs the framework developed here to test whether this is the case over real-world datasets [citation redacted].

In sum, to successfully model audit studies, one would have to make $\tau_{\text{gen}}$ metaphysically plausible as a view of what constitutes gender; define the effect of high-level interventions on gender (e.g.) as a geometric mean of the corresponding effects at the low-level; and define $D$ and $\epsilon$ so that this alignment is approximately causally consistent. As we have underscored, each of these tasks comes with associated empirical and normative challenges.

### 4.2.2 Normative relevance

Suppose, however, that these problems are addressed, and that we have defined the effects of gender as some complex function of the effects of its lower-level constituents. Then if the alignment is approximately causally consistent, we could rest assured that audit studies revealed an effect of gender, i.e. that we acquire information about (1) via (2):

$$\mathbb{P}[\mathsf{Interview\ Callback} \mid \mathrm{do}(\mathrm{Gen}_1)] - \mathbb{P}[\mathsf{Interview\ Callback} \mid \mathrm{do}(\mathrm{Gen}_2)] \tag{1}$$

$$\mathbb{P}[\mathsf{Interview\ Callback} \mid \mathrm{do}(\mathrm{Resume}_1)] - \mathbb{P}[\mathsf{Interview\ Callback} \mid \mathrm{do}(\mathrm{Resume}_2)]. \tag{2}$$

Even when this assumption is granted, there remains a second question about why the interventional difference (2) is normatively relevant. For instance, the populations corresponding to the resumes used in the audit study may be *atypical*, insofar as there is no guarantee that when one changes "Andrew" to "Anne" on a resume, one moves from a representative population of men to a representative population of women (Figure 5). If the resumes created by audit studies are highly atypical, it is unclear how they could shed light on a normatively salient population; the interventions performed by audit studies would simply be "strange" (Hu and Kohler-Hausmann, 2020, p. 10).

[11] It is an open question whether a metaphysically plausible alignment between gender and its constituents can deliver cross-contextual causal consistency, even approximately. See Tolbert (2024a, p. 1105) and Tolbert (2024b) for skepticism that the effects of race are homogenous enough to lend racial categories explanatory power; in the framework of abstraction, this could lead to skepticism that there exists a cross-contextual, causally consistent alignment between race and its constituents.

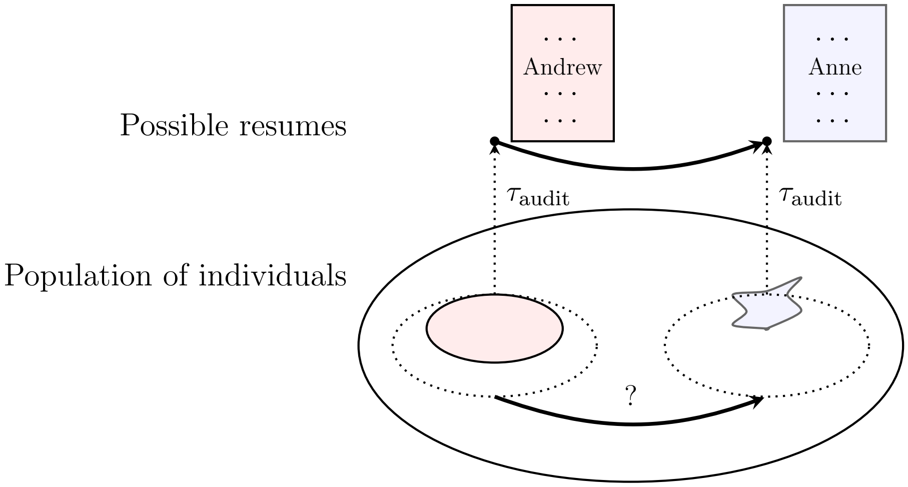


Figure 5: Atypicality in the resume audit study. The resumes differ only in names, but the "Andrew" resume corresponds to a large population of individuals (the red oval), who are representative of the larger group of people named Andrew (the leftmost dotted oval). By contrast, the "Anne" resume corresponds to a small, atypical population of individuals (the blue splotch), rather than the larger group of people named Anne (the rightmost dotted oval).

That said, normative relevance need not require typicality. For example, in the resume audit study, an auditor who believes that women students who study engineering should receive similar job opportunities as men might decide that the causal contrast between women and men engineers is relevant, even if the former population is "atypical" (i.e. significantly smaller). This choice of a causal contrast would be based on a normative assumption that these populations deserve similar treatment. Likewise, concluding from the *absence* of such an interventional difference that there is not discrimination requires a normative assumption that there are no other contrasts that are normatively relevant—for instance, contrasts corresponding to differences in the schools typically attended by women and men.

In sum, a successful model of audit studies of gender discrimination via names would require both an alignment between gender and its constituents which conveys information about high-level effects through low-level interventions on names (§4.2.1), and an argument that the populations corresponding to interventions on names are normatively relevant (§4.2.2). The assumption that this can be done may be subjected to further normative and empirical challenges, and we make no claim to have defended it. However, these further debates target particular modeling assumptions, which can be stated within the framework of causal abstraction; unlike the modularity problem, they raise no worries for causal models of discrimination as such. For example, the obstacle to modeling audit studies is not modularity, but it could be a failure of causal consistency. But causal consistency is a notion that is defined within the framework of causal abstraction.

## 4.3 Experimental design

We now illustrate how the above challenges affect experimental design. Imagine an in-person audit study, in which trained actors of different genders attend a job interview, presenting identical resumes and answering interview questions identically (Hu 2024a p. 8; cf. Kohler-Hausmann 2018, p. 1216). An experimenter must determine whether to match or experimentally vary the appearances and mannerisms of actors of different genders—for instance, whether they are wearing a dress or a suit, whether or not they wear makeup, or how assertive they are.

When audit studies are modeled using the Single-Level Assumption, experimental designs must always be licensed by bespoke modularity assumptions. Any features of individuals that experimenters vary are treated as "part of" gender, while any feature they keep constant is treated as "separate from" gender, and placed in the causal diagram subject to a modularity assumption. When a variable is modular with respect to gender, it cannot be partially constitutive of it. Thus, different choices of intervention correspond to different assumptions about how gender is socially constructed.

This tight connection between assumptions about the social construction of gender and an audit study's choice of an intervention can lead to implausible and normatively undesirable conclusions about the constitution of gender, even by the auditor's own lights. Suppose an auditor designs a causal test of gender discrimination looking at differences in treatment between gender-nonconforming men and gender-conforming women. To vary gender-conformity across applicants, both applicants wear skirts. This implies, on the Single-level Assumption, that dress is modular with respect to gender. But this modularity assumption is strange in light of an underlying motivation of the experiment, namely that there are highly gendered standards about dress, which would suggest that dress partially constitutes gender.

On the abstraction picture, by contrast, given the assumption that wearing typically feminine attire is partially constitutive of gender, we can understand the auditor's choice of intervention as a matter of selecting normatively relevant populations for similar treatment (§4.2.2). The auditor running a study about gender-nonconforming job applicants includes dress in their abstraction, corresponding to their understanding that standards of dress are gendered. They then select an intervention that compares women wearing skirts and men wearing skirts, corresponding to the normatively relevant comparison of treatment for gender-conforming women and gender-nonconforming men. Their choice of intervention is made *in light* of how dress constitutes gender, rather than by excluding dress from the constitution of gender.

Of course, in the abstraction framework, it is still possible for the crux of an intervention identification to rest on what attributes constitute gender. *Bostock vs. Clayton* provides one example. In this case, the plaintiff, a gay male employee, was fired for his sexual orientation. The majority decided that this amounted to sexual discrimination, since had the male employee been a female employee attracted to men, she would not have been fired. The dissenters replied that had the male employee been a female employee with same-sex attraction, she would have still been fired (Dembroff et al., 2020; Kohler-Hausmann and Dembroff, 2022). Should we vary the individual's sexual orientation to test for gender discrimination, or just vary biological sex features and hold orientation fixed?[12]

[12] We put aside the question of whether discrimination on the basis of sexual orientation can be understood as a category additionally meriting protection from discrimination apart from sex. Our more limited aim is to articulate what distinguishes two proposals for counterfactual tests of gender discrimination in the abstraction framework, one similar to that of Dembroff et al. (2020), and another one given in the dissenting opinion of SCOTUS (2020). We do not analyze the majority decision of Bostock precisely on its own terms

The former approach, which varies the individual's sex but not that of their partner, does not require a distinction between gender conformity and non-conformity in the model of gender, and indeed the dissents in SCOTUS (2020) are motivated by the view that effects of sexual orientation do not constitute effects of gender.[13] By contrast, on the latter approach, the high-level model in the abstraction might possess a single variable with four possible values: gender-conforming man, gender-conforming woman, gender-nonconforming man, and gender-nonconforming woman. The lower-level attribute of the plaintiff being same-sex attracted would then be aligned with being a gender-nonconforming man, so varying sexual orientation is a way of intervening on gender (from gender-conforming man to gender-nonconforming man). Many social constructionist theories of gender indeed intend to highlight norms punishing non-conformity (cf. Kohler-Hausmann and Dembroff 2022, pp. 88-89). Such modeling choices might be further justified, in part, by considerations of causal consistency; if gender-conforming and gender-nonconforming men are in fact treated very differently, then without such a high-level distinction, the abstraction will not be very causally consistent, combining groups that are treated very differently together.

In summary, the causal abstraction framework distinguishes a number of assumptions pivotal to the identification of an intervention. While the choice *to* vary a certain attribute commits us to including it in the constitutive basis of a social kind (as with sexual orientation in *Bostock vs. Clayton*), the choice *not to* experimentally vary an attribute need not be motivated by its exclusion from the constitutive basis of the social kind (as with attire in the in-person audit study). In other words, being included in the constitutive basis for the relevant social kind is a necessary condition for being varied in an experimental design, but not a sufficient one; a feature may be held fixed, in order to ensure that the test targets normatively relevant subpopulations.

# 5 Conclusion

While pervasive, the notion of causal discrimination is difficult to spell out in a precise and plausible way. Modularity requires that protected attributes like gender and race are separable from other attributes, and social constructionism casts doubt on this requirement. In response, many have proposed that we give up on modeling gender as causing worse treatment, either arguing that a distinct attribute (e.g. perception of gender) causes worse treatment, or that traditional causal models of discrimination are fundamentally misguided.

This paper has addressed the problem differently. We introduced a framework for reasoning about discrimination, in which gender is a high-level abstraction of lower-level, in-principle manipulable features. In this framework, gender can be modeled as itself causing worse treatment. The essential condition of modularity is ensured by allowing assumptions about social construction to be precisely and explicitly stated, via an alignment between

– the opinion takes sex as biological, and consequentially, as Dembroff et al. (2020), pp. 7-8, and Kohler-Hausmann and Dembroff (2022), pp. 84-86, point out, commits a modularity violation just like the dissent.

[13] For instance: "The Court tries to convince readers that it is merely enforcing the terms of the statute [prohibiting sex discrimination], but that is preposterous. Even as understood today, the concept of discrimination because of 'sex' is different from discrimination because of 'sexual orientation."' (SCOTUS (2020), dissent of Alito p. 3).

gender and its lower-level constituents. Such assumptions can then be subjected to further challenges. How do we define social kinds and their effects? Which population contrasts are normatively relevant to discrimination? These questions have an empirical aspect and a normative one: they recommend different experimental designs, and their answers will depend in part on what makes discrimination wrongful. While answers to these questions lie beyond the scope of this paper, we have used the abstraction framework to state them precisely, and to lay out some of their implications. We conclude that these questions point to important avenues for further work at the intersection between ethics and philosophy of science, but not to any in-principle objections to causal models of discrimination as such.

## Acknowledgements

Many thanks to Hannah DeBrine, Myra Deng, Johann Frick, Issa Kohler-Hausmann, Elek Lane, Russell McIntosh, Christian Nakazawa, Naftali Weinberger, and Eliza Wells for comments and discussion, and to Caltech's Linde Center for Science, Society, and Policy (LCSSP) for supporting a research incubator to develop this paper. We are also grateful for helpful comments from audiences at the University of California, Berkeley, Emory University, the University of Michigan, the University of Pittsburgh, and Stanford University. Special thanks to two anonymous reviewers from Noûs for extremely detailed and helpful feedback. Kara Schechtman was supported by a graduate fellowship award from Knight-Hennessy Scholars at Stanford University during initial work on this paper.

# References

Ásta (2013). The social construction of human kinds. *Hypatia*, 28(4):716–732.

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In Geffner, H., Dechter, R., and Halpern, J., editors, *Probabilistic and Causal Inference: the Works of Judea Pearl*, pages 507–556. ACM Books.

Barnes, E. (2018). Symmetric dependence. In Bliss, R. and Priest, G., editors, *Reality and Its Structure: Essays in Fundamentality*, pages 50–69. Oxford University Press, Oxford, UK.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Beckers, S., Eberhardt, F., and Halpern, J. Y. (2020). Approximate causal abstractions. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, pages 606–615.

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685. Number: 01.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Working Paper 9873, National Bureau of Economic Research.

Bright, L. K., Malinsky, D., and Thompson, M. (2016). Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81.

Cherry, F. and Bendick Jr., M. (2018). Making it Count: Discrimination Auditing and the Activist Scholar Tradition. In Gaddis, S. M., editor, *Audit Studies: Behind the Scenes with Theory, Method and Nuance*, volume 14 of *Methodos*, pages 45–62. Springer.

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1):139–167.

Dembroff, R., Kohler-Hausmann, I., and Sugarman, E. (2020). What Taylor Swift and Beyoncé teach us about sex and causes. *U. Pa. L. Rev. Online*, 169:1.

Gaddis, S. M. (2018). *An introduction to audit studies in the social sciences.* Springer.

Gaebler, J. D., Goel, S., Huq, A., and Tambe, P. (2024). Auditing large language models for race & gender disparities: Implications for artificial intelligence-based hiring. *Behavioral Science & Policy*, 10(2):46–55.

Galos, D. R. and Coppock, A. (2023). Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments. *Science Advances*, 9(18):eade7979.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research.*

Glymour, C. and Glymour, M. R. (2014). Commentary: Race and sex are causes. *Epidemiology*, 25(4):488–490.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–85.

Haslanger, S. (2017). Gender and social construction: Who? what? when? where? how? In *Applied Ethics*, pages 299–306. Routledge.

Hellman, D. (2008). *When is discrimination wrong?* Harvard University Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hu, L. (2024a). Normative Facts and Causal Structure. *The Journal of Philosophy*, forthcoming.

Hu, L. (2024b). What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, pages 1–26.

Hu, L. and Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency.*

Hu, L. and Kohler-Hausmann, I. (2024). What is Perceived When Race is Perceived and Why It Matters for Causal Inference and Discrimination Studies. *Law & Society Review*, forthcoming.

Janzing, D. and Mejia, S. H. G. (2022). Phenomenological causality.

Janzing, D. and Schölkopf, B. (2008). Causal inference using the algorithmic Markov condition. arXiv:0804.3678 [math].

Jenkins, K. (2016). Amelioration and inclusion: Gender identity and the concept of woman. *Ethics*, 126(2):394–421.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems (NIPS)*, 30.

Kohler-Hausmann, I. (2018). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113(5):1163–1228.

Kohler-Hausmann, I. and Dembroff, R. (2022). Supreme Confusion About Causality at the Supreme Court. *City University of New York Law Review*, 25(1).

Kolodny, N. (2023). *The Pecking Order: Social Hierarchy as a Philosophical Problem*. Harvard University Press, Cambridge, Massachusetts.

Kusner, M., Russell, C., Loftus, J., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems (NIPS)*.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17):556–567. Seventieth Annual Meeting of the American Philosophical Association Eastern Division.

Loi, M., Nappo, F., and Viganò, E. (2023). How I would have been differently treated. discrimination through the lens of counterfactual fairness. *Res Publica*, 29(2):185–211.

Mallon, R. (2024). Naturalistic Approaches to Social Construction. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition.

Marcellesi, A. (2013). Is race a cause? *Philosophy of Science*, 80(5):650–659.

Narayanan, A. (2022). The limits of the quantitative approach to discrimination. James Baldwin Lecture [transcript], Princeton University.

Pearl, J. (2000). *Causality*. Cambridge University Press, New York.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA.

Plečko, D. and Bareinboim, E. (2024). Causal fairness analysis. *Foundations and Trends in Machine Learning*, 17(3):304–589.

Rischel, E. F. and Weichwald, S. (2021). Compositional abstraction error and a category of causal models. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR. ISSN: 2640-3498.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI).*

SCOTUS (2020). Bostock v. Clayton County. Supreme Court of the United States. Docket No. 17-1618.

Sen, M. and Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Singh, K. and Wodak, D. (2023). Does race best explain racial discrimination? *Philosophers' Imprint*, 23.

Spirtes, P. and Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5):833–845.

Thomsen, F. K. (2018). Direct discrimination. In Lippert-Rasmussen, K., editor, *Routledge Handbook of Discrimination*, pages 19–29. Routledge.

Tolbert, A. (2024a). Causal agnosticism about race: Variable selection problems in causal inference. *Philosophy of Science*, pages 1–11.

Tolbert, A. W. (2024b). Restricted racial realism: Heterogeneous effects and the instability of race. *Philosophy of the Social Sciences.*

VanderWeele, T. J. and Robinson, W. R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484.

Wang, A., Ramaswamy, V. V., and Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Weinberger, N. (2022). Signal manipulation and the causal analysis of racial discrimination. *Ergo.*

Weslake, B. (2024). Exclusion excluded. In Wilson, A. and Robertson, K., editors, *Levels of Explanation*, pages 101–135. Oxford University Press.

Woodward, J. (2005). *Making things happen: A theory of causal explanation.* Oxford university press.

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research*, 91(2):303–347.

Woodward, J. (2021). Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese*, 198(1):237–265.

Woodward, J. and Hitchcock, C. (2003a). Explanatory generalizations, part i: A counterfactual account. *Nous*, 37:1–24.

Woodward, J. and Hitchcock, C. (2003b). Explanatory generalizations, part ii: Plumbing explanatory depth. *Nous*, 37:181–199.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, 101(2):245–280.