The Use and Misuse of Counterfactuals in Ethical Machine Learning

Atoosa Kasirzadeh
University of Toronto
Australian National University
atoosa.kasirzadeh@anu.edu.au

Andrew Smart
Google
andrewsmart@google.com

ABSTRACT

The use of counterfactuals for considerations of algorithmic fairness and explainability is gaining prominence within the machine learning community and industry. This paper argues for more caution with the use of counterfactuals when the facts to be considered are social categories such as race or gender. We review a broad body of papers from philosophy and social sciences on social ontology and the semantics of counterfactuals, and we conclude that the counterfactual approach in machine learning fairness and social explainability can require an incoherent theory of what social categories are. Our findings suggest that most often the social categories may not admit counterfactual manipulation, and hence may not appropriately satisfy the demands for evaluating the truth or falsity of counterfactuals. This is important because the widespread use of counterfactuals in machine learning can lead to misleading results when applied in high-stakes domains. Accordingly, we argue that even though counterfactuals play an essential part in some causal inferences, their use for questions of algorithmic fairness and social explanations can create more problems than they resolve. Our positive result is a set of tenets about using counterfactuals for fairness and explanations in machine learning.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; **Machine learning**; • **Social and professional topics** → **Socio-technical systems**; **Race and ethnicity**.

KEYWORDS

Ethics of AI, Ethical AI, Counterfactuals, Machine learning, Fairness, Algorithmic Fairness, Explanation, Explainable AI, Philosophy, Social ontology, Social category, Social kind, Philosophy of AI

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445886

ACM Reference Format:
Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3442188.3445886

1 INTRODUCTION

The use of counterfactuals has become increasingly popular in the machine learning community for many reasons such as making sense of algorithmic fairness or explainability in automated decision-making for consequential social contexts [4, 9, 14, 18, 29, 35, 42, 47, 51]. As a result, machine learning algorithms coupled with counterfactuals could be used for making high-stakes decisions with ethical and legal impacts in domains such as insurance, predictive policing, and hiring. Despite this widespread attention and use, there is a surprising lack of engagement with the long-standing philosophical and social scientific literature on the required ontological and semantic conditions for an appropriate application of counterfactuals.

What is a counterfactual? Consider X and Y to represent events or facts and the following chain of occurrences "X and Y", where X precedes Y in time. A counterfactual analysis can help to find whether X is a cause of Y by supposing the non-occurrence of X and seeking for the effect of this supposition on Y. This corresponds to evaluating whether the counterfactual 'If X had not occurred, Y would not have occurred.' is true. In machine learning practice, there are several technical ways to generate and evaluate counterfactuals, such as feature-based explanations, prototype explanations, example-based explanations, or causal explanations [19, 32, 35, 38, 46, 50, 51]. These approaches are most often rooted, implicitly or explicitly, in either of the two prominent conceptual approaches for evaluating counterfactuals: the close-enough-possible-worlds approach inspired by Lewis [36] and Stalnaker [49], and the causal modeling approach developed by Spirtes et al. [48] and Pearl [44], among others.[1]

To evaluate a counterfactual, the close-enough-possible-worlds approach compares the actual world in which X and Y occur with those similar-enough worlds to the actual world in which X does not occur (e.g., comparing a data instance

[1]Strictly speaking, [36, 49] develop the closest-possible-worlds approach to make sense of counterfactuals. With a bit of weakening, the (set of) closest-possible-world(s) can be interpreted as the (set of) close-enough-possible-world(s), where due to practical considerations those possible worlds that are close enough to the actual world (rather than the closest possible worlds) are selected. For a recent alternative to evaluating conditionals relative to a causal model see [1].

to a similar data instance or to a prototype when generating example-based or prototype explanations, respectively, requires comparison with respect to some notion of enough similarity). If in those worlds Y does not occur the counterfactual is considered true and X is deemed the cause of Y; otherwise, the counterfactual is deemed false. The close-enough-possible-worlds account has been mainly used in discussions of counterfactual explanations in machine learning and the causal modeling approach has been widely applied for examining fairness counterfactually. Although these two semantic accounts are very different, the following abstract recipe is common to both for the evaluation of counterfactuals. First, determine the facts to be kept fixed under counterfactual variation. Second, vary the antecedent. Third, determine the influence of the variation on the consequent.

In this paper, we explore the ontological and epistemological-semantic conditions required for using either of the two conceptual approaches for an appropriate application of counterfactuals to ethical machine learning, in particular to algorithmic fairness and social explanations. We argue that in some cases, the lack of a right grounding of the elements of a counterfactual into the social world can lead to their misuse in machine learning applications. We review a broad body of papers from philosophy and social sciences on the ontology of social categories and conclude that the counterfactual approach in machine learning fairness and social explainability might require an incoherent theory of what some social categories such as race are. Our findings suggest that despite its appeal for convenient analysis of fairness and social explanations, most often the social categories may not admit an apt counterfactual intervention, and hence may not appropriately satisfy the required assumptions for evaluating the truth or falsity of counterfactuals. Accordingly, we argue that even though counterfactuals play an essential part in some causal inferences, their use in discussions of algorithmic fairness and social explanations can create more problems than they resolve.

Related work and novelty. Before we go further, we would like to explicitly contrast our paper in more detail with related work to highlight its novelty. There are four main closely related works on this topic which explicitly or implicitly critique counterfactual theories of social causation in decision-making contexts. Kohler-Hausmann [34] argues that the counterfactual causal model is *wrong* for detecting discrimination in both law and social science. Building on this idea, Hu and Kohler-Hausmann [30] argue that perhaps we need to use a formal model other than causal models (such as constitutive diagrams) for detecting discrimination. Hanna et al. [23] use critical race theory and argue that the multi-dimensionality of race should be taken into account whenever this phenomenon becomes relevant to the machine learning community, and challenges practitioners to explicitly ask who is doing the categorizing and for what purpose? Barocas et al. [4] discuss the mapping of the explanatory features to actions in the world when using feature-highlighting explanations. We share the perspective of these authors. However, the novelty of our contribution is threefold. (1) We provide a conceptual analysis of the vagueness of the notion of 'similarity', rooted in the close-enough-possible-worlds approach. This approach is the conceptual basis of feature-based, prototype, and example-based analytic methods for examining counterfactuals by machine learning community. The notion of similarity is used in almost all conceptions of counterfactual explanations or fairness as referenced. To the best of our knowledge, the philosophical-conceptual basis [36, 37, 49] and assumptions required to assess the 'similarity' of counterfactual worlds/scenarios are not properly examined in the machine learning literature, yet 'similarity' is used, implicitly or explicitly, for making sense of counterfactual explanations or fairness. (2) We go beyond the mere criticism of causal modeling as applied to the social domain, and consider counterfactuals more generally by examining both the close-enough-possible-worlds account and causal modeling. We think that just a critique of manipulating social categories is not sufficient because in disciplines such as medicine and public health, the use of protected attributes such as race or gender are considered to be an ethically acceptable component of research (e.g., prostate cancer screening [5, 21]).[2] (3) We provide positive results in terms of a set of detailed tenets as summarized in table 1, showing that any trace of a counterfactually fair or explainable algorithm (in a social context) involves making several choices and value judgments. To that end, the implicit presumptions, choices, and value judgments must be made as explicit and obvious as possible by using table 1. No related work does (1) – (3).

The rest of the paper is structured as follows. In Section 2, we examine the two prominent approaches to modeling and evaluating counterfactuals, the close-enough-possible-worlds and the causal modeling approaches, in more detail. In Section 3, we discuss the use of counterfactuals for analyzing fairness and social explanations in machine learning practice before raising ontological and epistemological-semantic problems from this use in Section 4. In Section 5, we suggest a set of tenets about the use of counterfactuals in machine learning. Section 6 concludes the paper.

2 BACKGROUND: THE CLOSE-ENOUGH-POSSIBLE-WORLDS AND CAUSAL MODELING

Consider the following counterfactuals: (1) If Suzy had not thrown the rock, the window would not have shattered. (2) If Nora had not been Latina, she would not have been denied admission. Are these counterfactuals true or false? Does 'Suzy's throwing the rock' cause 'the shattering of the window'? Does 'Nora's being Latina' cause 'denying admission'? There are two prominent approaches to evaluate counterfactuals, the close-enough-possible-worlds approach that is mainly used in the discussions of social counterfactual explanations [51],

[2] We are not promoting this use. We just report that in medicine, economics, public health and other related disciplines, the use of protected classes such as race or gender sometimes is the basis of development or allocation of some resources.

and the causal modeling approach that is at the center of discussions about counterfactual fairness [35].[3] We present these two semantic approaches independently, though we must mention that, theoretically speaking, the relationship between the two is not that straightforward [7]. For the lack of space, we cannot go into the differential details in this paper. But we translate this lack of straightforward connection between the two semantic approaches into our set of principles for using counterfactuals in machine learning research.

According to the closest-possible-worlds view [36, 49], a counterfactual can be treated syntactically and semantically via a variant of a modal logic for counterfactuals. The evaluation of the counterfactual X $\Box\!\!\rightarrow$ Y (if X had occurred, Y would have occurred) requires the specification of a set of possible worlds in which X occurs. If in these possible worlds Y also occurs, the counterfactual X $\Box\!\!\rightarrow$ Y is true. These possible worlds must be ordered in terms of comparative similarity or closeness to the actual world (in which X occurs and Y occurs). For instance, if in all the worlds which are close enough to the actual world except that Suzy does not throw the rock, the window does not shatter, then Suzy's throw is the cause of the shattering of the window. If in all the close-enough-possible-worlds to the actual world in which Nora is not Latina, she is not denied admission, then Nora's being Latina is the cause of her rejection. The close-enough-possible-worlds approach to the evaluation of counterfactuals requires an ordering of the possible worlds in terms of similarity to the actual world. In Section 4, we discuss that the notion of *similarity* is inherently vague and that the similarity ordering can be done in many different ways. As a result, depending on *the choices* for the similarity criteria and the ordering, we can obtain contradictory judgments about the truth or falsity of counterfactuals. Hence, the vagueness and the multiplicity of orderings pertain to the problems of using counterfactuals in machine learning.

A causal modeling approach uses a causal model as a representational tool for exploring the space of alternative causal hypothesis. Following Pearl [44], from a causal modeling perspective, the world is described in terms of random variables and their values. The random variables are either exogenous or endogenous, and they might take continuous or categorical values. The exogenous variables ($\mathcal{U}$) are determined by factors outside of the causal model, and serve as fixed background assumptions to the causal reasoning. The endogenous variables ($\mathcal{V}$) may have a causal influence on each other. This influence is modeled by a set of structural equations $\mathcal{F}$ that are functions for capturing the potential causal effects of functional dependencies on the endogenous variables. A set of exogenous and endogenous variables, their values, and a set of structural equations form a causal model $\mathcal{M}$=($\mathcal{U}, \mathcal{V}, \mathcal{F}$). $\mathcal{M}$ can be graphically visualized by a directed acyclic graph. This graph facilitates cognitive efforts in thinking about potential causal sources, effects, and causal relations. In such a graph, a node represents a random variable and an edge between each pair of nodes represents a direct causal relation between the corresponding random variables; for instance, X is a direct cause (parent) of Y is represented by $X \rightarrow Y$. Nodes with no incoming edge are said to be exogenous.

To find causal relations via a causal model requires establishing well-defined connections between some aspects of the sample data and a causal model [44, 48]. The main connections are often captured by two causal assumptions, the causal Markov condition and faithfulness. The causal Markov condition ensures that a variable is independent of its non-descendants given its parents. The causal faithfulness condition requires that all inter-dependencies in the observational data are non-accidental and structural, the result of the structure of the causal graph. To counterfactually think via a causal modeling approach in a specific machine learning domain requires an in-depth interpretation of the mapping of the random variables on the elements of the domain and the satisfaction of the causal assumptions. If the domain of counterfactual thinking occurs at the level of the social world, we require an apt interpretation of the mapping of the random variables on social categories, the relationship between them, and the meaning of causal assumptions applied to the relevant categories. So far, we have provided a discussion of the two most prominent semantic approaches to the evaluation of counterfactuals. In the next section, we give two examples of the use of counterfactuals in machine learning: in understanding fairness (via causal modeling) [35] and in understanding social explanations (via the closest-possible-worlds) [51].

3 COUNTERFACTUALS IN ETHICAL MACHINE LEARNING

3.1 Counterfactual fairness

Discussions about the treatment of *fairness* in machine learning systems have primarily taken place in relation to a group or the individual level. To achieve group fairness, a (statistical) measure must compare a predictor's behavior across different protected demographic groups, and then seeks for approximate parity of some desirable statistical measure across the groups [8, 25]. On the other hand, a measure of individual fairness must compare a predictor's behavior across similar individuals [11, 31]. To date, the most popular proposal for making sense of individual fairness has been the use of causal modeling for interpreting individual fairness in a counterfactual way [35]. Kusner et al. [35] define a fair predictor to be the one that gives the same prediction had the individual were different, for example, had the individual been of another race or gender. This demands an implicit assumption that other features and properties (except for the tweaked category in the causal model) remain the same for that individual. More precisely, Kusner et al. (2017) gives the following definition: counterfactual fairness "captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group."

Consider a prediction-based problem characterized in terms of A (a set of protected attributes), X (a set of non-protected attributes), and Y (the prediction output). To put this problem

[3] Kilbertus et al. [33] use causal models to analyze fairness. We focus our discussion on Kusner et al. [35], but our criticism also applies to their work.

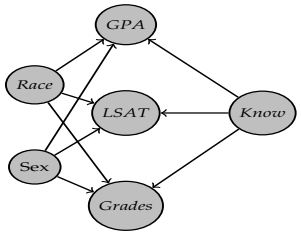

Figure 1: A causal model for a fair predictor adapted from [35].

into a causal modeling schema requires fixing U, the set of exogenous variables. Following [35], the definition of counterfactual fairness for the predictor $\hat{Y}$ stipulates the satisfaction of the following condition for $X{=}x$ and $A{=}a$, for all y, and any value a' attainable by A:

$P(\hat{Y}_{A\leftarrow a}(U){=}y|X{=}x, A{=}a){=}P(\hat{Y}_{A\leftarrow a'}(U){=}y|X{=}x, A{=}a')$

To make matters more concrete, we focus on an example of a machine learning system, as discussed by [35], employing a predictor $\hat{Y}$ to decide who should be admitted to law school based on its prediction of potential student's first year grade (figure 1). The algorithm makes the prediction according to knowledge about the following attributes of individuals: gender, race, GPA, and law school entrance exam (LSAT). According to [35], the set of sensitive attributes are A {sex, race}, and the non-sensitive ones are X {GPA, law school entrance exam}. Moreover, there is a causal link set between the attributes and the prediction of potential student's first year grade. To make this classifier fair, the following question should be answered: what would the predictor have predicted, if the individual had a different race (a different sensitive attribute)? This use of counterfactuals requires assuming a single change (race) or a limited set of changes (such as sex and race) to an individual, and then evaluate the probabilistic condition above given the supposition that everything else remains the same for that individual. Although the proposal might sound simple, in the next section we discuss the problems pertaining to this proposal such as requiring commitment to a peculiar conception of race as well as controversial views about the integrity of what an individual (or the perception of an individual) is, for the purpose of satisfying the convenient requirements of counterfactual modeling and evaluation. Or, to understand what counterfactual fairness is, we first need to make choices about which counterfactual worlds to consider and the basis by which the closeness of counterfactual worlds (including the knowledge of how a counterfactually different version of the target individuals) to the actual world is specified.

3.2 Social counterfactual explanation

Counterfactual explanations are claimed to be among the most popular types of explanations for opaque algorithmic decisions [51]. For instance, let us assume Nora has applied for a mortgage and her application is denied via an algorithmic system. A counterfactual explanation for this denial can be: If Nora's annual income had been $60,000, she would have received the loan. As a matter of fact, Nora is denied a loan and her annual income is $40,000. Or consider the following counterfactual explanation: If Nora had not been Latina, she would not have been denied the loan. As a matter of fact, Nora is denied a loan and she is Latina. This is an instance that requires making a putative plausible assumption about a different version of Nora (with only a different race, everything else equal to the original version of Nora), and then trust the validity of this explanation.

In the next section, we offer two main arguments challenging a counterfactual approach to algorithmic fairness and social explanations when the things that require counterfactual supposition are social categories such as gender, sexual orientation, or race in terms of which (the uniqueness of) a person is characterized.

4 TWO PROBLEMS

In this section, we specify two sets of problems, ontological and epistemological-semantic, that one faces upon attempting to construct a fair or explainable classifier which incorporates the counterfactual supposition of social categories such as race and gender, as sketched in Section 3. The problems arise in the attempts to answer the following questions: what are the objects of manipulation? Which counterfactual worlds are similar enough to the actual world or whose causal model's perspective should we care about?

4.1 What is manipulated?

There has been a long standing debate among several disciplines such as philosophy, sociology, law and epidemiology about the causal effects of social categories such as race and gender [12, 16, 34, 41]. To counterfactually suppose a social attribute *of an individual* requires first specifying what the social categories are and what it means to suppose a different version of an individual with the counterfactually manipulated social property. This counterfactual question amounts to asking, "what if person X had not been "race Y" or "gender Z"?

There are several competing contemporary schools of thought about what social categories are, and our review here is merely representative of some and by no means exhaustive. In the rest of this section, we take 'race' as a prototypical instance of the social categories of interest to counterfactual manipulation. With some modifications, similar arguments can be made about other social categories such as gender.

Roughly, we can distinguish between three major positions about what race is [39, 40]. The geo-biological essentialism about race largely signifies dividing humans into a sufficiently small, discrete number of categories, usually for the purposes of colonial conquest, enslavement or domination of one group over another [10]. The categorization has been based on some kind of biological foundation (e.g., modern genes) essential to humans, and inherited from one generation to another. This conception of race identifies some geo-biological features (such as skin color, hair texture, and eye form) that are only

common to the members of a racial group, usually from a specific geographical region. The geo-biological conception of race has been questioned extensively, and has been critically challenged by scientific and philosophical arguments ranging from denying that the concept of race has any biological foundations to denying the very existence of races. Some also have argued that this biologically essentialist view about race cannot be separated from the political project of racial oppression, domination and disenfranchisement [10, 23]. In addition to the geo-biological ancestry conception of race, there are two other major views about the ontology of race.

On the one hand, racial skeptics argue for the falsity of naturalism about race and conclude that no type of race exists [2, 3, 20, 52–54]. They claim that the natural candidates for the bases of race such as geography, phenotypes, and geneology fail according to scientific findings. The normative implications of this ontological view is to entirely disregard the existence of race. On the other hand, racial constructivists dismiss the conception of biological race, but argue that the concept of race must be preserved for the purpose of social movements and affirmative action to abolish social and structural injustice. How so? One of the most influential proponents of racial constructivism, Haslanger [15, 26, 27], suggests a group-based understanding of race marked by ancestry and appearance and by hierarchical relations of power *for the purposes of fighting against social injustice*. This conception of race finds using 'race' as a justifiable entity for the purpose of resisting and combating racism. Other than that, racial identification by the dominant group constrains the autonomy of individuals by requiring them to be what a specific racial group signifies from the point of view of who has defined it. Social constructivism hence maintains that a social category – be it racial, gender, or class – was brought into existence or shaped by historical events, social forces, political power, or colonial conquest, all of which could have been very different [6, 12, 20]. Being a social constructivist about race and gender means that one does not subscribe to the view that race and gender are natural or biological categories with permanent or immutable properties. In other words, for such a constructivist, the term 'race' cannot refer to an essentially biological attribute such as skin tone, a genetically produced trait, or a signifier that people just have and thereby obviously belong to a designated racial group [34].

Kusner et al. [35] claim that it is counterproductive to assume social categories such as race cannot be causes because we can design experiments on such categories by intervening on a particular aspect of the attribute 'race', such as 'race perception'. We disagree. We think this claim only serves to justify the convenient assumptions required for causal modeling (i.e., that conception of race is amenable to counterfactual manipulation). As we have shown above, there is no universally agreed-upon perception of race. To be able to talk about the causal effect of social categories, we first need to specify what these categories are. For instance, we might be justified in first having a robust social ontology informed by critical theory [28]. Only after this exploration, we are able to discuss what our perception of race is. As we have seen, there is a plurality of responses to this question, and our response depends on the perspective we adopt about this matter.

Recall that an algorithm that subscribes to counterfactual fairness requires evaluating the actual non-occurrence of X with the supposition that X did occur. For example, we should replace the actual person (or our perception thereof) who has a protected attribute, such as being Latina, with a counterfactual version of the same person who has a different protected attribute, such as being white, to test whether the algorithm makes the same prediction about the actual person (or our perception thereof) and the counterfactual person. What view about race (or perception of race) does it require to suppose that racial category non-Latina for the counterfactual version of person i knowing that Latina is the real feature of person i? Counterfactual fairness (or counterfactual social explanation) requires us to force a random variable to take a certain value. Is the required counterfactual suppositions for designing a fair algorithm compatible with the views about race specified above?

Racial skepticism is ruled out as an alternative of commitments held by the proponents of counterfactual fairness or counterfactual social explanations due to its denial of the very existence of such categories. Social constructivism makes sense of race *for the purposes of fighting against social injustice*. Hence, the constructivist ontology of race has, in addition, a purpose-relative reality that the algorithm must reflect in its reasoning and arguably is not subject to counterfactual variation separate from the scope of the fight against social injustice. Perhaps the only viable theory of race that remains for counterfactual fairness requires commitment to a reductionist view about social categories such as race or gender as biological attributes. Several scholars have argued that this commitment is deeply problematic (see, for instance, [34]). We share this perspective for several decision contexts. This purely reductionist understanding of social categories as essential and physical attributes, in addition to being scientifically outdated, fails the task of robust objectivity, and might indirectly widen and exaggerate the problematic associations between the sensitive attributes that are the result of social and structural injustice in the first place.

4.2 Similarity between worlds and the view from somewhere

Is there an objective view from nowhere form which to assess the validity of counterfactuals? In this section, we raise some epistemological-semantic problems for comparing and selecting the set of counterfactual possible worlds that are close enough to the actual world.

First, we focus on the problem of inherent vagueness associated with similarity between possible worlds. Counterfactual scenarios in counterfactual worlds stand in contrast to actual scenarios in actual worlds. To evaluate a counterfactual requires a comparison between an actual world and a set of sufficiently similar counterfactual worlds to the actual world. The counterfactual $X \mathrel{\square\!\!\rightarrow} Y$ is true just in case it takes less of a departure from the actual world to make X true along with

Y than to make X true without Y. But, which counterfactual worlds? The number of counterfactual worlds is myriad (perhaps even uncountably infinite). Lewis and Stalnaker [36, 49] emphasize that the counterfactual worlds of interest to the actual world are the ones that are the most similar to the actual world. In some cases of comparing natural features between worlds, it is possible to arrive at a consensus for the ordering of similar worlds. However, in many cases the vagueness of this notion is problematic and counterintuitive for the evaluation of counterfactuals [17]. Lewis [37] provides some guidance to ordering possible worlds: (1) avoid big widespread violations of the laws of nature of the actual world, (2) maximize the spatiotemporal perfect match of particular matters of fact, (3) avoid small, localized violations of the laws of nature of the actual world, and (4) secure approximate similarity of particular matters of fact. But, how to translate these considerations to the social domain? Further research is required to understand how to avoid big widespread violations of commitments to our ontological views about social categories in the possible worlds framework.

The ordering of similar worlds faces severe problems because for some ordinary counterfactuals, some irrelevant possible worlds end up determining the counterfactuals' truth values. Also, depending on what kind of possible worlds we choose, we might end up assigning a different truth-value to a counterfactual statement. To make the matters more concrete, consider the following counterfactual [13] (3) If Nixon had pressed the button, then there would have been a nuclear holocaust. A similarity-based approach requires the following truth-evaluation: (3) is true if and only if the worlds most similar to the actual world in which Nixon pressed the button, there was a nuclear holocaust. But the worlds in which there is a nuclear holocaust are drastically different from the actual world: the entire future history of humanity would be different in such a world. This example points to the difficulties we face in making judgments about the ordering of possible worlds.

The causal modeling approach for interpreting counterfactuals builds on Lewis's ordering of similar worlds. However, it appeals to the cognitive architecture of the human mind in order to resolve the arbitrariness of assumptions about the ordering of the counterfactual worlds. Pearl [45] argues that to make sense of the notion of "similarity" we should rely on the fact that we experience the same world and share the same mental model of its causal structure. However, relying on a largely speculative psychological theory of how the human mind handles the infinity of possible counterfactual worlds does not resolve the normative and ethical implications of choosing which possible worlds are the most similar to the actual world.

Indeed, different epistemic view points might suggest different ordering of possible worlds. After all, humans differ extensively in the standpoints from which they observe the world, and these standpoints influence the formation of causal mental models [24]. From an abstract point of view, a causal model is specified according to a set of nodes, edges, and assumptions. What these nodes and edges represent and how they are interpreted suggest a particular standpoint about the organization of world from the view point of the causal model. The crucial point to remember is that no causal model captures absolutely objective relations in the world. Depending on the convenient assumptions for a causal model, X can be counterfactually dependent on Y in one model but not in another [22]. These convenient assumptions specifying the causal model might enforce some false perceptions about the social world (at the risk of being seriously wrong). This suggests that there is always a view from somewhere, as opposed to a more objective and universal "view from nowhere" [43] from which we can assess whether a counterfactual is assertible.

5 RESULTS OF OUR ANALYSIS

So far, we have argued that the use of counterfactuals in fair and explainable machine learning is not straightforward, and that there are various trade-offs and value judgments essential to the use of counterfactuals for ethical machine learning. Examination of all these assumptions produces awareness about various trade-offs, value judgments, or potential harms of using or misusing counterfactuals. Therefore, to aptly use counterfactuals requires bringing forth all implicit and unspecified assumptions about the ontology of the categories on which we run counterfactual analysis as well as the epistemic and the interpretational issues pertaining to the evaluation of counterfactuals. Examination of all these assumptions produces awareness about some unexpected potential harms that can result from the laudable goals of fair and explainable machine learning.

In this section, we offer strategies for specifying and reflecting on the hidden ontological and epistemological-semantic assumptions through an interdisciplinary conversation. We summarize the results of our study (Table 1) by suggesting a detailed set of tenets to check and reflect upon before applying counterfactuals to fair and explainable machine learning. Following this set of tenets would enable modellers and algorithmic designers to state unspecified and implicit assumptions about social ontology as explicitly as possible. It also suggests a path to researchers for seeking a variety of justifications in seeing the social world through a counterfactual lens, and to become aware of some potential harms and disadvantages of making sense of fairness and explanations counterfactually. Our results are a necessary step to perform before designing and applying some putative counterfactually fair or explainable algorithms to social contexts.

Table 1 has three columns. The first column provides a category of different kinds of presumptions and choices (ontological and epistemological-semantic) which are necessary to examine before designing and applying counterfactually fair and explainable algorithms. The second column provides the set of questions to ask and answer for articulating the implicit set of assumptions in column one as explicit as possible. The third column gives an exemplar of the questions to answer in the context of a particular social problem.

Recall the counterfactual "If Nora had not been Latina, she would not have been denied admission."

Assumption	Question	Example
Ontological perspective	What are the social categories?	What is race (or gender)?
Ontological choice	What ontological perspective do we choose to adopt, and why?	Among different views, what do we take race to be? Social constructivism? Geo-biological ancestry conception of race? why?
Ontological knowledge	How do we know about the social categories?	Who do we consult about the conception of race?
Semantic choice	Close-enough-possible-worlds or causal modeling? What is the justification?	Why do we choose either of these semantic approaches to counterfactually suppose that Nora is not Latina? How is our choice justified?
Evaluation reliability	What happens to the truth value of the counterfactuals of interest if we change the semantic approach? How robust is the truth value of the counterfactual when moving from a close-enough-possible-worlds approach to causal modeling?	Is the truth value for "If Nora had not been Latina, she would not have been denied admission." differ when we choose the semantic approach?
Similarity choice	How do we choose what similarity means in this context?	What do we sacrifice by supposing a particular cluster of similar worlds (rather than other possible clusters of similar worlds) in which an individual is the same except for their race?
Comparison criteria	What are our chosen criteria for comparing the similar worlds of interests to the actual world? Are these criteria socially warranted?	What characterization for comparing similar worlds justifies keeping (almost) everything about a person fixed except for their race? What does this socially mean?
Idealization	What do we miss by translating social categories into random variables?	What is left out by translating an individual's race to a random variable?
Context	How do these categories operate in the world?	How does race function in the world? Does this conflict with the assumptions necessary for counterfactual manipulation of race?
Ethical and social harm	Does our ontological preference generate harms in relation to social justice (combating structural injustices)?	Does our ontological preference for what race is generate harms in relation to combating racial injustice?

Table 1: Any use of a counterfactually fair or explainable algorithm (in a social context) involves making several ontological, semantic and ethical choices and judgments. These implicit presumptions, choices, and judgments must be made as explicit and obvious as possible.

Here are the ontological assumptions that should become explicit. First, an explicit statement of the ontological perspective the algorithmic system is adopting. To tweak "being Latina" the designers of the system need to specify what race (e.g., Latina) from their perspective is. Second, the designers could discover whether a counterfactual approach inadvertently commits them to a problematic social ontology. They could provide morally and politically appropriate justifications for why, among other options, they choose and adopt this ontological perspective about race. Is it because this conception of race is compatible with some simplistic assumptions about social ontology that are required to use a causal modeling approach? What are the genuine reasons for this choice, in relation to respecting intellectual humility for what we know about race from other disciplines? Third, the assumptions about ontological knowledge should become explicit. For instance, who do we consult about a theory of race, and why?

The epistemological-semantic presumptions and choices that must be made explicit are as follows. First, what is the semantic choice? Will we choose a close-enough-possible-worlds or causal modeling approach? What is the justification for this choice? Does our choice make a difference to the truth evaluation of the counterfactual for this particular context of employment such as Nora not being Latina? Second, how do we account for the evaluation of reliability? How robust is the truth value of the counterfactual when moving from a close-enough-possible-worlds approach to causal modeling? For instance, is the truth value of "If Nora had not been Latina, she would not have been denied admission." differ when we choose either of the semantic approaches? Third, how do we decide about the meaning of similarity in the particular context of employment? What do we sacrifice by supposing a particular cluster of similar worlds (rather than other possible cluster of similar worlds) in which an individual is the same except for their race? Or if we are specifying that everything that is not causally dependent on the tweaked category should remain constant, how do we know what is not causally dependent? Fourth, what are our chosen criteria for comparing the possible similar worlds of interest to the actual world? Are these criteria socially warranted? For instance, what characterization for comparing similar worlds justifies keeping almost everything about a person fixed except for their race? What does this socially mean? Fifth, there are questions about the translation of social categories such as race into random variables that can be appropriately treated by an algorithm, if the semantic choice is causal modeling. What is left out if we translate the conception of race into random variables? Does that matter? Why or why not? Sixth, there are questions about the choice of context. How do social categories (such as race) operate in the world? Does this conflict with the required assumptions for counterfactual manipulation of race? Finally, there are questions about some ethical harms that can result from the use of counterfactual analysis. Does our ontological preference generate harms in relation to some desired social justice agenda? For example, does our ontological preference for what race is generate harms in relation to some affirmative action plans for combating racial injustice?

In sum, Table 1 shows that any trace of a counterfactually fair or explainable algorithm (in a social context) involves making several choices and presumptions. By following these tenets, computer scientists can discuss the validity and the implications of these choices in accordance with other disciplines such as philosophy, social sciences, and anthropology. To that end, the implicit presumptions and choices will be made as explicit and obvious as possible, and an interdisciplinary conversation can result in concluding whether the counterfactuals should be used in the generation of explanations and fairness in machine learning practice.

6 CONCLUSION

Counterfactuals are increasingly applied in machine learning, for example in designing fair and explainable algorithms. This paper provides a detailed set of principles, according to philosophical and social scientific insights, for articulating the implicit and unspecified contextual presumptions and choices made in counterfactual applications. Regardless of which evaluation approach to counterfactuals one takes, this set of principles could help researchers to conduct interdisciplinary conversations and become aware of the potential harms and ethical impacts of their counterfactual thinking as it pertains to the social world. We think this set of principles is an example of how to establish a successful interdisciplinary conversation between machine learning researchers and social scientists, philosophers, and ethicists.

7 ACKNOWLEDGEMENTS

We would like to thank Alex Beutel, Yoni Halpern, Manasi Joshi, Christina Greer, Robert Williamson, Mario Günther, and members of the Humanizing Intelligence Grand Challenge at Australian National University for extremely helpful comments and feedback. We would also like to thank participants in the Workshop on Philosophy and Medical AI at the University of Tübingen, NeurIPS's Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability, and the Bias and Fairness in AI Workshop in Ghent, Belgium for critical discussion.

REFERENCES

[1] Holger Andreas and Mario Günther. 2020. A Ramsey Test analysis of causation for causal models. *The British Journal for the Philosophy of Science* (2020).
[2] Anthony Appiah. 1995. The uncompleted argument: Du Bois and the illusion of race. *Critical inquiry* 12, 1 (1995), 21–37.
[3] Anthony Appiah. 1996. Race, Culture, Identity: Misunderstood Connections." Color Conscious: the political morality of race. Anthony Appiah and Amy Gutmann.
[4] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.
[5] Yoav Ben-Shlomo, Simon Evans, Fowzia Ibrahim, Biral Patel, Ken Anson, Frank Chinegwundoh, Cathy Corbishley, Danny Dorling, Bethan Thomas, David Gillatt, et al. 2008. The risk of prostate cancer amongst black men in the United Kingdom: the PROCESS cohort study. *European Urology* 53, 1 (2008), 99–105.

[6] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons.

[7] R. Briggs. 2012. Interventionist counterfactuals. *Philosophical studies* 160, 1 (2012), 139–166.

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[9] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 582–593.

[10] Lory Dance. 2010. Struggles of the Disenfranchised: Commonalities Among Native Americans, Black Americans, and Palestinians. *Al-Hewar Magazine* (2010).

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[12] Dave Elder-Vass. 2012. Towards a realist social constructionism. *Sociologia, problemas e práticas* 70 (2012), 9–24.

[13] Kit Fine. 1975. Critical notice. *Mind* 84, 335 (1975), 451–458.

[14] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. Measuring Social Biases of Crowd Workers using Counterfactual Queries. *arXiv preprint arXiv:2004.02028* (2020).

[15] Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. 2019. *What is Race?: Four Philosophical Views*. Oxford University Press.

[16] Clark Glymour and Madelyn R Glymour. 2014. Commentary: race and sex are causes. *Epidemiology* 25, 4 (2014), 488–490.

[17] Nelson Goodman. 1972. Seven strictures on similarity. In *Problems and projects*. New York, Bobbs-Merrill.

[18] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245* (2018).

[19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[20] Ian Hacking, Jan Hacking, et al. 1999. *The social construction of what?* Harvard university press.

[21] Susan Halabi, Sandipan Dutta, Catherine M Tangen, Mark Rosenthal, Daniel P Petrylak, Ian M Thompson Jr, Kim N Chi, John C Araujo, Christopher Logothetis, David I Quinn, et al. 2019. Overall survival of black and white men with metastatic castration-resistant prostate cancer treated with docetaxel. *Journal of Clinical Oncology* 37, 5 (2019), 403.

[22] Joseph Y Halpern and Christopher Hitchcock. 2011. Actual causation and the art of modeling. *arXiv preprint arXiv:1106.2652* (2011).

[23] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.

[24] Sandra G Harding. 2004. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press.

[25] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[26] Sally Haslanger. 2000. Gender and race:(What) are they?(What) do we want them to be? *Noûs* 34, 1 (2000), 31–55.

[27] Sally Haslanger. 2010. Language, politics, and "the folk": looking for "the meaning" of 'race'. *The Monist* 93, 2 (2010), 169–187.

[28] Sally Haslanger. 2016. What is a (social) structural explanation? *Philosophical Studies* 173, 1 (2016), 113–130.

[29] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809* (2018).

[30] Lily Hu and Issa Kohler-Hausmann. 2020. What's Sex Got To Do With Machine Learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[31] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

[32] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).

[33] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[34] Issa Kohler-Hausmann. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.* 113 (2018), 1163.

[35] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[36] David Lewis. 1973. *Counterfactuals*. Oxford: Blackwell.

[37] David Lewis. 1986. *Philosophical papers II*. Oxford: Oxford University Press.

[38] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

[39] Ron Mallon. 2004. Passing, traveling and reality: Social constructionism and the metaphysics of race. *Noûs* 38, 4 (2004), 644–673.

[40] Ron Mallon. 2007. A field guide to social construction. *Philosophy Compass* 2, 1 (2007), 93–108.

[41] Alexandre Marcellesi. 2013. Is race a cause? *Philosophy of Science* 80, 5 (2013), 650–659.

[42] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[43] Thomas Nagel. 1989. *The view from nowhere*. Oxford University Press.

[44] Judea Pearl. 2009. *Causality*. Second Edition, Cambridge University Press.

[45] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[47] Kacper Sokol and Peter A Flach. 2018. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant.. In *IJCAI*. 5868–5870.

[48] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[49] Robert C Stalnaker. 1968. A theory of conditionals. In *Studies in Logical Theory*, Nicholas Rescher (Ed.). Oxford: Blackwell, 98–112.

[50] Arnaud Van Looveren and Janis Klaise. 2019. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584* (2019).

[51] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2017), 841.

[52] Jennifer K Wagner, Joon-Ho Yu, Jayne O Ifekwunigwe, Tanya M Harrell, Michael J Bamshad, and Charmaine D Royal. 2017. Anthropologists' views on race, ancestry, and genetics. *American Journal of Physical Anthropology* 162, 2 (2017), 318–327.

[53] Naomi Zack. 1994. *Race and mixed race*. Temple University Press.

[54] Naomi Zack. 2014. *Philosophy of science and race*. Routledge.