Variation semantics: when counterfactuals in explanations of algorithmic decisions are true[*]

Laurenz Hudetz   Neil Crawford

14 March 2022

Abstract

We propose a new semantics for counterfactual conditionals. It is primarily motivated by the need for an adequate framework for evaluating counterfactual explanations of algorithmic decisions. We argue that orthodox Lewis-Stalnaker similarity semantics and interventionist causal modelling semantics are not adequate frameworks because they classify too many counterfactuals as true. Our proposed semantics overcomes this problem of the orthodox approaches and has further advantages, including simplicity, robustness, closeness to practice and applicability. It is based on the idea that a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true at an elementary possibility ω just in case C is true at all variants of ω at which A is true, other things being equal. We provide a novel explication of the idea of a variation that makes a given sentence true while leaving other things (but not necessarily *all* other things) equal.

[*] Neil Crawford and Laurenz Hudetz are jointly responsible for sections 1 and 2 and general editing. Laurenz Hudetz is responsible for sections 3-7.

Contents

1 Introduction

Algorithms are increasingly used for decision-making in many areas, including criminal justice, finance, health care, education and hiring.[1] For example, there are algorithms that calculate an estimate of the probability that a prisoner will re-offend based on available data about the prisoner as input. Decisions about whether to grant or deny parole are being made in light of the outputs of such algorithms (Wexler, 2017). In many cases, explanations of algorithmic decisions are desirable or even required.[2] For example, a prisoner who has been denied parole based on the prediction of an algorithm arguably deserves an explanation of why they have been denied parole. This raises the question what counts as an explanation in such contexts and when a given explanation is good.

We focus on a prominent type of explanations, namely *counterfactual explanations* (Wachter et al., 2018). To illustrate, suppose a person applies for a loan but their application is rejected based on the output of an algorithm calculating credit scores. Then a counterfactual explanation could look like this:

> Your loan application was rejected because your income was £$38,000$. If your income had been at least £$40,000$, your credit score would have been above the threshold.

There can be multiple counterfactual explanations of the same fact. But any counterfactual explanation involves a counterfactual conditional (for short: counterfactual) — a sentence of the form 'if A were the case, C would be the case' or 'if A had been the case, C would have been the case' (in symbols: $A \mathrel{\Box\!\!\rightarrow} C$).[3] The antecedent A is typically a statement regarding one or more input variables or variables that can be calculated in terms of input variables, and the consequent C is typically a statement regarding the output variable(s) of the algorithm used for making the decision.

[1] Roughly, an algorithm is a mechanical method for determining an output given an input of appropriate type. More formally, we view algorithms as computable functions. We restrict attention to deterministic algorithms in this paper. Nondeterministic algorithms can return different outputs for the same input.

[2] There is an ongoing legal debate about whether the European Union's General Data Protection Regulation (GDPR) contains a right to explanation (see Kaminski, 2018; Wachter et al., 2017), and there are philosophical arguments for a moral to explanation (Vredenburgh, 2022).

[3] We ignore differences in tense and use the formalisation $A \mathrel{\Box\!\!\rightarrow} C$ for conditionals both of type II (past simple) and type III (past perfect).

Counterfactual explanations can be useful for several purposes (Wachter et al., 2018). They give data subjects conditions that would lead or would have led to a particular outcome. This can help data subjects to understand decisions (Wachter et al., 2018), and it can enable them to make changes to achieve a desired outcome (Dandl et al., 2020). But even if an attribute is difficult or impossible to change, a counterfactual involving it can still be useful. For instance, counterfactuals can reveal biases of algorithms (Wachter et al., 2018). If a change of a protected characteristic in the input data would lead to a different outcome, this can be an indicator of bias. Counterfactual explanations can also help to discover inaccuracies in the input data, for instance when the antecedent of the conditional used in the explanation is in fact satisfied by the data subject. (A concrete example will be given in section 2.1.) Counterfactual explanations can thus enable data subjects to exercise their right to contest decisions based on illegitimate discrimination or inaccurate or incomplete data (Wachter et al., 2018).

To assess how good a given counterfactual explanation is, two questions regarding the conditional on which the explanation is based need to be addressed: (a) Is the conditional true? (b) Is the conditional useful? That the conditional is true is a minimal requirement for any good counterfactual explanation in most contexts. But truth is not sufficient for usefulness. Whether a true counterfactual is useful typically depends on its purpose and various pragmatic and contextual factors. For instance, suppose the purpose of a counterfactual is to enable a data subject to make changes that lead to a desired outcome. It is widely held that, to be useful in that case, the antecedent should describe only minimal changes to a small number of attributes that are relatively easy to modify for the data subject, and it should not mention changes to any protected characteristics (Verma et al., 2020, 2021). However, if its purpose is to reveal biases or other flaws of an algorithm, changes to protected characteristics should not be ruled out, and changes need not be minimal or easy to bring about. So, standards of usefulness are best chosen on a case-by-case basis. However, there should be a general and rigorous approach to the question concerning truth.

Although truth conditions for counterfactual conditionals are of fundamental importance for assessing counterfactual explanations, they have not yet been explicitly addressed in the machine learning literature on counterfactual explanations.[4] So far this literature has been focussing on developing methods for finding possible input data for which an algorithm returns desired outcomes and which are "close" to the actual input data.[5] Such methods yield lists of possible input data along with the associated outputs of the algorithm. To illustrate, here is a simplified example of such a list due to Molnar (2019, section 9.3). It concerns a machine learning model estimating the probability that a customer's credit risk is classified as 'good'. Values in bold font are those that are changed relative to the original input data in order to achieve an outcome above 0.5.

age	sex	job	amount	duration	outcome
58	f	**skilled**	6143	**28**	**0.501**
55	**m**	unskilled	6143	**24**	**0.501**
54	**m**	unskilled	**4889**	**24**	**0.506**
28	**m**	**skilled**	6143	**24**	**0.590**

Methods for generating such lists of counterfactual data points are undoubtedly important for *finding* useful counterfactual conditionals. But this type of work leaves open under which conditions counterfactual conditionals are true. In particular, it leaves open the truth conditions of counterfactuals that specify ranges of values (rather than precise point values) or that have logically complex parts, e.g. sentences like this: 'The outcome would be between 0.5 and 0.7 if (a) the credit amount were between £4000 and £4600, or (b) the duration were between 12 and 18 months, or (c) the credit amount were between £4800 and £5200, and the duration were between 18 and 24.'

But counterfactuals of that type are important for good counterfactual explanations. They are typically more informative and easier to understand for a data subject than a list of individual data points with associated outputs. For instance, given just a list of data points, it is unclear how robust the associated outputs of the algorithm are under perturbations of the inputs. It is unclear how close the values of a data subject's variables need to be to a data point on the list to yield a particular outcome. Ranges may therefore often be more useful than precise

[4] For recent literature reviews see Artelt and Hammer (2019), Stepin et al. (2021), and Verma et al. (2020).

[5] Examples of such methods include the method of Multi-Objective Counterfactuals (MOC) by Dandl et al. (2020), the method of Diverse Counterfactual Explanations (DiCE) by Mothilal et al. (2019), the Recourse method by Ustun et al. (2019) and the Feature Tweaking method by Tolomei et al. (2017).

point values. And disjunctive antecedents help to summarise alternative ways of reaching an outcome.

To determine the truth values of counterfactuals that have logically complex parts or involve ranges, a rigorous semantics with general truth conditions is needed. One such semantics is Lewis-Stalnaker similarity semantics, which provides truth conditions in terms of comparative similarity relations between possible worlds. Similarity semantics is not only one of the most prominent (if not *the* dominant) account of counterfactuals in philosophy, it is also similar in spirit to machine learning approaches to counterfactual explanations that use similarity metrics to find suitable "nearby" variants of the actual input data.

However, similarity semantics has a substantial problem: it classifies some counterfactuals about the behaviour of algorithms as true that should arguably be classified as false (as we show in section 2). Interventionist semantics (Briggs, 2012; Pearl, 2009), which is a main alternative to similarity semantics enjoying popularity in computer science, has a similar problem (as we argue in section 4.2). Additional concerns about similarity semantics and interventionist semantics have led some philosophers and machine learning researchers to call for caution with the use of counterfactual explanations (Kasirzadeh & Smart, 2021). There are concerns regarding (a) interventionist manipulations of social categories such as race and gender and (b) difficulties in making judgements about overall similarity between worlds. However, we believe that these concerns do not warrant rejecting counterfactual explanations per se. These concerns rest on the assumption that counterfactuals are evaluated either in a similarity-based or interventionist framework. But we reject this assumption. We propose an alternative semantics that overcomes the above-mentioned problems, called 'variation semantics' (section 3).

Variation semantics is based on the intuitive idea that a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true at a given elementary possibility ω iff C is true at all variants of ω where A is true, other things being equal. We explicate the idea of a variation that makes a given sentence true while leaving "other things equal" in a novel way. On our account, "leaving other things equal" does not entail "leaving *all* other things equal". Our explication is neither based on similarity comparisons nor on surgical interventions. Instead, it focuses on upstream changes to relevant independent variables.

We discuss important properties of variation semantics in sections 4–6. We outline its advantages compared to some of the most prominent semantics for counterfactuals: similarity semantics, interventionist causal modelling semantics and Kit Fine's abstract truth-maker semantics. We argue that variation semantics is more suitable for evaluating counterfactuals in explanations of algorithmic decisions than its alternatives. But its scope of applicability is not limited to such counterfactuals. It is much more general than it might seem at first glance. We illustrate this using an example. Finally, we comment on philosophical implications as well as implications for practice (section 7).

2 Challenges for similarity semantics

To motivate the truth conditions of variation semantics, it is instructive to first examine some problems of similarity semantics. The key idea of similarity semantics is to take into account how similar ("close") different possible worlds are to each other when determining whether a given counterfactual is true (see Lewis, 1973a; Lewis, 1973b; Stalnaker, 1968). In similarity semantics, a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is true at a possible world w iff some world where A and C are both true is closer to w than any world where A is true but C is false (if there are any worlds where A is true). In case there is a *closest* world where A is true, the truth condition reduces to: $A \mathrel{\Box\!\!\rightarrow} C$ is true at w iff C is true at all the closest worlds to w where A is true.

When using similarity semantics to evaluate counterfactuals about the behaviour of algorithms, it is not necessary to understand 'possible worlds' in a modal realist or any other metaphysically substantive sense. Instead, one can can take possible data points as the points at which sentences are evaluated. Possible data points constitute the kind of elementary possibilities that is relevant when reasoning about alternative inputs and outputs of an algorithm.[6] Moreover, one can define comparative similarity orderings in terms of state-of-the-art measures of similarity or distance between data points. Data scientists have proposed and discussed various similarity metrics in the context of counterfactual explanations.[7]Which counterfactuals are classified as true in similarity semantics depends

[6] Roughly, a possible data point is an assignment of values to the variables relevant to the algorithm that satisfies the presuppositions of the algorithm (if any) and any integrity constraints of the underlying database.

[7] In the context of counterfactual explanations, two important proposals are weighted Man-

on the choice of a particular similarity ordering (or metric), and this choice is beset with difficulties (Fine, 1975; Kasirzadeh & Smart, 2021).

But there is another problem with similarity semantics. Given mild assumptions about ‘comparative overall similarity’, it yields implausible results when it comes to certain counterfactuals that have logically complex parts or that involve ranges of values. There are issues with non-monotonic relationships, ranges in antecedents, and disjunctive antecedents. Let us examine these three issues in more detail.

2.1 Non-monotonic relationships

Similarity semantics yields problematic results when it comes to machine learning models where the output variable depends non-monotonically on some input variable. To illustrate this issue, let us take the algorithm QCovid as an example. QCovid calculates an estimate of a person’s risk of death from COVID-19 based on their health records (Clift et al., 2020). One of the variables used by QCovid is BMI, which is a predictor of risk of death from COVID-19. It is important to note that risk of death depends non-monotonically on BMI. A very high BMI increases the risk compared to a medium BMI. But a very low BMI also increases the risk compared to a medium BMI.

QCovid made headlines in March 2021 because many people were incorrectly classified as being at high risk. As reported by *The Guardian*, “if a patient’s weight or ethnicity are not recorded on their health records, QCovid automatically ascribes them a BMI of 31 (obese) and the highest risk ethnicity (black African)” (Geddes, 2021). Counterfactual explanations can aid trouble-shooting in cases like this. For instance, a counterfactual of the following form could be useful:

(C1) If your BMI were in the interval $[t_1, t_2]$, your QCovid risk score would not be high.

Suppose you are informed that your QCovid risk score is high, and suppose you are given a counterfactual explanation based on C1. Suppose C1 is true but your actual BMI lies in the specified interval $[t_1, t_2]$. Then this indicates an inaccuracy in your health record. And if it turns out that your BMI is missing in your health

hattan metrics (Wachter et al., 2018) and the Gower distance (Dandl et al., 2020).

record, the truth of C1 indicates an issue concerning the variable BMI in the algorithm.

It is crucial that counterfactuals are true when used for purposes like this, and an adequate approach to determining their truth values is needed. But similarity semantics is not adequate. It classifies some sentences of the form of C1 as true although they should come out as false, namely when the interval $[t_1, t_2]$ is too large, containing very low BMI values that result in a high risk score.

For instance, let ω_0 be a patient's available health data, and suppose BMI is inaccurately set to 31 in ω_0 although the patient's BMI is much lower in reality. Suppose the algorithm returns a high risk score for all input data that are like ω_0 except that BMI is strictly greater than t_2. Recall that t_2 is the upper threshold mentioned in C1. We assume that 31 lies above that threshold. So ω_0 gets a high risk score. Suppose further that the algorithm also outputs a high risk score for all input data that are like ω_0 except that BMI is below some threshold t_1^*. Let t_1^* be above the lower threshold t_1 mentioned in C1 (but still below t_2). Finally, suppose the algorithm returns a risk score that is not high for all those input data that are like ω_0 except that BMI lies in the interval $[t_1^*, t_2]$. Figure 1 illustrates this scenario.

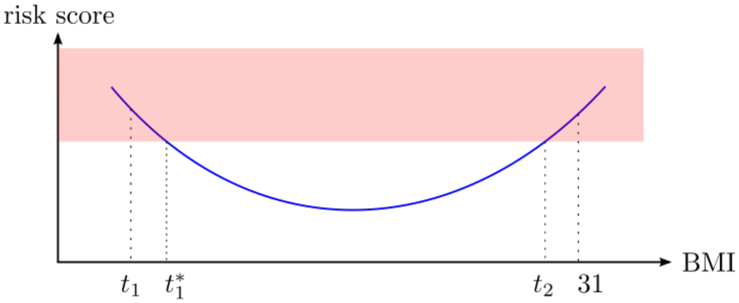

Figure 1: Risk score depending on BMI. Scores in the shaded region are high.

Given these assumptions, the algorithm returns a high risk score for all those inputs that are like ω_0 except that BMI lies in the interval $[t_1, t_1^*)$, which is a subset of the interval $[t_1, t_2]$ specified in C1. In this scenario, C1 should in our view be classified as false at ω_0.[8] Having a BMI in the interval $[t_1, t_2]$, other things being

[8] However, note that the might-counterfactual "If your BMI were in the interval $[t_1, t_2]$, your

equal, is not sufficient for not having a high risk score. If the patient is underweight with an actual BMI in the subinterval $[t_1, t_1^*)$, their risk score is high also in light of the corrected data.

But according to similarity semantics, C1 is true in this scenario. Under any reasonable construal of ‘comparative overall similarity’, the closest possible data point to ω_0 satisfying the antecedent of C1 is the data point where BMI is set to t_2 but all other input variables have the same values as in ω_0. Given this data point as input, the algorithm outputs a risk score that is not high. So the consequent is true at the *closest* point where the antecedent is true. Thus, C1 is true at ω_0 according to similarity semantics although it should come out as false.

Note that if risk of death from COVID-19 depended *monotonically* on BMI (so, the lower the BMI, the lower the risk), there would be no problem in this scenario. For then we could conclude from the fact that the risk score is not high at the closest point where the antecedent is true (i.e. where BMI $= t_2$) that it is also not high at any other points where BMI is in the interval $[t_1, t_2]$, other things being equal. So, similarity semantics would get the truth value right.

However, one should not conclude from this that non-monotonic models are the problem here rather than similarity semantics. First, although monotonicity is sometimes a desideratum for models in machine learning, non-monotonic models cannot be brushed aside. Some important relationships simply are non-monotonic, and accurate models will reflect that. Second, there is a more general underlying problem that cannot be solved by discarding non-monotonic models, namely the problem of ranges.

2.2 Ranges

Similarity semantics has a general problem with counterfactuals that specify ranges of values in the antecedent — even when it comes to monotonic models. To illustrate this, consider a very simple algorithm for classifying LDL (low-density lipoproteins) cholesterol levels:

risk score *might* not be high” is arguably true at ω_0. After all, the consequent is true in some possible data points where BMI lies in the specified interval, other things being equal.

LDL level (input)	classification (output)
less than 100 mg/dL	optimal
100 to 129 mg/dL	near optimal
130 to 159 mg/dL	borderline high
160 or higher	high

The relationship between the numerical input and the ordinal output is monotonic. The higher the LDL level, the higher the classification.[9] Nevertheless, similarity semantics yields implausible truth values for counterfactuals regarding this algorithm. For example, suppose a nutritionist makes the following claim when advising a patient:

(C2) If your LDL cholesterol level were in the range from 100 to 200 mg/dL, it would be classified as 'near optimal'.

This claim seems false. For it entails, on a pre-theoretic understanding, that even high LDL values of up to 200 mg/dL would be classified as 'near optimal'. But this is not the case.

Contrary to this intuitive judgement, similarity semantics classifies C2 as true, provided that the patient's LDL level lies below 100 mg/dL. In that case, the closest possibility that makes the antecedent of C2 true is having an LDL level of 100 mg/dL. This level is classified as 'near optimal'. So, the consequent is true at the *closest* point at which the antecedent is true. Clearly, there are also many possible ways in which the antecedent of C2 can be true while the consequent is false. But similarity semantics considers only the closest possibilities where the antecedent is true. All other possibilities are neglected. And therefore similarity semantics has a general problem when it comes to counterfactuals specifying ranges of values.

2.3 Disjunctive antecedents

The problem of disjunctive antecedents is another, related issue. It a well-known objection to similarity semantics that goes back to Fine (1975). Here is an example that illustrates the problem and its relevance to explanations of algorithmic decisions. Suppose you apply for a loan. You feed the application form with data

[9] Nothing depends on the fact that there are four classification ranges in this example. The problem arises whenever there are at least two.

including your annual income (£39,000) and your age (60 years). You are denied the loan because the bank's algorithm returns a credit score below a certain threshold t when fed with your data. Now, suppose that in this situation the following counterfactual conditional is true:

(C3) If you were 40 years old or your annual income were £40,000, your credit score would be above t.

Suppose you made a mistake when entering your data and your actual age is 40 years. So your corrected data actually satisfy the antecedent of C3. Then it seems reasonable to infer that the credit score based on your corrected data must be above the threshold if C3 is true. So you should be successful when making use of your right to contest decisions based on inaccurate data. After all, it seems plausible in this context that C3 logically entails the following simpler conditional:

(C4) If you were 40 years old, your credit score would be above t.

However, according to similarity semantics, C3 does not entail C4. If C3 is true in similarity semantics, the credit score based on your corrected data may still be below the threshold. Let us take a closer look at how this can be the case. Consider three data points:

(ω_0) the original data (age = 60, income = $38k$, ...),

(ω_1) the data point where income = $40k$ but all other input variables have the same values as in the original data,

(ω_2) the corrected data, i.e. the data point where age = 40 but all other input variables have the same values as in the original data.

Suppose the algorithm returns credit scores below t for ω_0 and ω_2 but a credit score above t for ω_1. So, even for the corrected data, your credit score remains below the threshold. Now, to apply the truth conditions of similarity semantics, we need some information about the comparative similarity ordering of possible data points. Suppose the ordering satisfies the following plausible conditions. First, data point ω_1 is closer to ω_0 than is ω_2. Earning a bit more is arguably less of a departure from actuality than being 20 years younger. Second, relative to ω_0, ω_1 is the closest data point where the antecedent of C3 is true. Setting the variable *income* to $40k$ is the smallest change one can make to satisfy the disjunction ‘age $= 40 \ \lor\ $ income $= 40k$’. Third, ω_2 is the closest data point, relative to ω_0, where the antecedent of C4 is true. So, setting the variable *age* to 40 is the smallest change one can make to satisfy ‘age = 40’.

In this scenario, C3 is true at ω_0 according to similarity semantics because the consequent of C3 is true at ω_1, which is the closest point at which the antecedent of C3 is true. In contrast, C4 is false at ω_0 according to similarity semantics. This is so because its consequent is false at ω_2, which is the closest point where the antecedent of C4 is true. But this means that C3 does not entail C4 according to similarity semantics. So, the credit score based on your corrected data can fall below the threshold even though C3 comes out as true in similarity semantics.

The root of the problem lies in classifying C3 as true at ω_0. This is highly misleading for the data subject and, in our view, simply incorrect. Our aim is to capture this intuition by providing an alternative semantics in which C3 does come out as false in this scenario and in which C3 entails C4. We acknowledge that there are also pragmatic approaches to the problem regarding disjunctive antecedents.[10] But a semantic solution is possible too, one that also solves the problems regarding ranges and non-monotonic models.

3 Variation semantics

The examples above suggest in our view that one should consider not only the closest data points where the antecedent is true when evaluating a counterfactual conditional. Instead, one should consider all ways in which the antecedent can be made true by changing relevant variables while holding everything else fixed. This is the core idea of variation semantics in a nutshell. For example, in our

[10] One pragmatic approach would be to insist that the “true” logical form of C3 is not that of a counterfactual with a disjuntive antecedent ($(A \lor B) \mathrel{\Box\!\!\rightarrow} C$) but rather a conjunction of two counterfactuals ($(A \mathrel{\Box\!\!\rightarrow} C) \land (B \mathrel{\Box\!\!\rightarrow} C)$). But this solution leads to bizarre results when carried over to the closely related problem of ranges. A counterfactual referring to a range in the antecedent would have to be understood as an infinite conjunction of uncountably many individual counterfactuals, one for each real number in the interval. Another pragmatic approach would be to claim that C3 is indeed true but it should not be asserted in this context. We see two problems with this. (1) The standard move would be say that C3 should not be asserted because there is a logically stronger (more informative) sentence one can assert instead. But this does not work here. The conditional ‘If your annual income were £40,000, your credit score would be above t’ is not logically stronger than C3 in similarity semantics. (2) It is unclear how the claim that C3 is true in this scenario could be justified without already presupposing similarity semantics.

example about COVID-19 risk above, one should consider all ways of making the antecedent 'BMI $\in [t_1, t_2]$' true by appropriately varying the value of the variable BMI in the given data point. If the consequent ('risk score is not high') comes out as true under all those variations, the counterfactual conditional is true, otherwise it is false.

To turn this idea into a rigorous semantic framework, we need to define several new concepts. The central concept is that of an *A-variant* of a possible data point, where A is a sentence. Roughly speaking, an A-variant of a possible data point ω is a possible data point at which the sentence A is true but that "otherwise" coincides with ω. Once this concept is made precise, the truth conditions of variation semantics can be stated in a simple form: a would-counterfactual 'if A were the case, C would be the case' ($A \mathrel{\Box\!\!\rightarrow} C$) is true at a data point ω iff C is true at every A-variant of ω. And a might-counterfactual 'if A were the case, C might be the case' ($A \mathrel{\Diamond\!\!\rightarrow} C$) is true at ω iff C is true at some A-variant of ω.

The rest of this section is devoted to explicating the concept of A-variants. To explicate this and related concepts, we need to define the notion of a *model* in variation semantics. But before presenting the technical definition, it is important to distinguish two senses of the word 'model' that must not be confused with each other: the scientist's sense and the logician's sense. By 'models in the scientists sense' we mean mathematical models based on formulas that establish functional relationships between dependent and independent variables. Mathematicians, scientists and engineers frequently understand 'model' in this way. Machine learning models are examples of models in this sense, but so are more traditional mathematical models from the natural and social sciences. By 'models in the logician's sense' we mean structures that allow us to assign truth values to sentences. It is this sense of 'model' that we have to make precise in variation semantics. For brevity, we call models in the scientist's sense simply 'mathematical models', and we call models in the logician's sense 'semantic models'.

When it comes to evaluating counterfactuals about the behaviour of mathematical models (call this the '*central intended application*' of variation semantics), there is a close relationship between semantic models and mathematical models. To assign truth values to sentences about the behaviour of a mathematical model, one needs a semantic model that is suitably related to the mathematical model in question. We therefore define semantic models in a way that allows us to incorporate mathematical models in semantic models. As usual in modal logic, semantic models are built on frames. But our notion of frames differs from traditional notions of frames. Traditionally, a frame is a set of possible worlds that is equipped with a relation between worlds (e.g. an accessibility relation or a similarity relation). In variation semantics, the structure of frames is chosen in a way that makes it straightforward to encode the central aspects of any given mathematical model M, namely (a) its dependent and independent variables and (b) the set of data points that are possible according to M, which captures the relationships between the variables in M.

Definition 1. (Ω, Var) is a **frame** in variation semantics iff

1. Ω is a non-empty set (the elements of which are called 'possible data points').
2. Var is a non-empty set of variables on Ω that is partitioned into independent variables (Var_i) and dependent variables (Var_d), i.e. each $X \in \mathit{Var}$ is a function from Ω to some value space $\mathcal{V}_X$, and either $X \in \mathit{Var}_i$ or $X \in \mathit{Var}_d$.
3. Possible data points are characterised by the values variables take in them, i.e. for all $\omega, \omega' \in \Omega$: $\omega = \omega'$ iff for all $X \in \mathit{Var}$, $X(\omega) = X(\omega')$.
4. Independent variables can be varied independently of each other, i.e. for every independent variable $X \in \mathit{Var}_i$, every possible value $v \in \mathcal{V}_X$ of X and every possible data point $\omega \in \Omega$, there exists a unique possible data point ω' at which X takes the value v but all other independent variables take the same values as in ω. (We denote this data point by '$\omega_{X:v}$'.)
5. The values of dependent variables are determined by (supervene on) the values of independent variables, i.e. any possible data points $\omega, \omega' \in \Omega$ that agree regarding the values of all variables in Var_i also agree regarding the values of all variables in Var_d.[11]

Some remarks about possible data points and variables are in order. Possible data points are taken as primitive for the sake of generality. If we defined possible data points, this would only make the semantics less flexible and decrease its range of applicability. But when it comes to the central intended application of variation

[11] In more formal terms: for all $\omega, \omega' \in \Omega$, if $X(\omega) = X(\omega')$ for all $X \in \mathit{Var}_i$, then $Y(\omega) = Y(\omega')$ for all $Y \in \mathit{Var}_d$.

semantics, we can say more about how possible data points and variables can be understood.

In the central intended application, we can take the label ‘possible data points’ seriously. But we understand them not as data points from available samples but rather as data points that are possible according to the mathematical model itself. Intuitively, a possible data point can be thought of as an assignment of values to *all* variables in *Var* (see condition 3) satisfying the following necessary and jointly sufficient conditions: (a) It must satisfy all assumptions of the mathematical model. For example, assigning the value 0 to a variable is not possible if this would lead to a division by zero. (b) It must satisfy the relationship between the dependent and independent variables postulated by the mathematical model. (c) It must satisfy any integrity constraints of the model’s underlying database if there is one. So, we can think of Ω as including a data point for each possible input to the mathematical model (see condition 4 above).

Variables are understood, essentially as in probability theory, as functions from Ω to value spaces. There are no constraints on the value spaces of variables.[12] Both qualitative variables (e.g. *gender*) and quantitative variables (e.g. *income*) can be handled by variation semantics. By ‘independent variables’ we mean the input variables of the mathematical model in question, also known as ‘features’. The dependent variables include at least the output variable(s), sometimes called the ‘target’; but intermediate variables may also be included.

To illustrate our definition of ‘frame’, let us specify a frame that captures a simple linear model with n independent variables $X_1, \ldots, X_n$ and one dependent variable Y related by the equation

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n,$$

where $\beta_0, \ldots, \beta_n$ are concrete real numbers. We must specify two things: the set of possible data points (Ω) and the set of variables (Var). First, let Ω be the set of all $(n+1)$-tuples $(x_1, \ldots, x_n, y)$ of real numbers such that $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$. These tuples are taken as the data points that are possible according to the simple linear model under consideration. Second, let $\mathit{Var} = \{X_1, \ldots, X_n, Y\}$, $\mathit{Var}_i = \{X_1, \ldots, X_n\}$ and $\mathit{Var}_d = \{Y\}$. Let all these variables be real-valued. That is,

[12] We do not even need the assumption that value spaces are measurable spaces.

every member of *Var* is a function from Ω to $\mathbb{R}$. For any $\omega = (x_1, \ldots, x_n, y) \in \Omega$, let $X_i(\omega) = x_i$ (for $1 \leq i \leq n$), and let $Y(\omega) = y$. Given this set-up, it follows that (Ω, *Var*) meets all conditions in definition 1 and that the linear model's equation is satisfied: for all $\omega \in \Omega$, $Y(\omega) = \beta_0 + \beta_1 X_1(\omega) + \ldots + \beta_n X_n(\omega)$.

This illustrates how to construct frames in the central intended application. But it is important to note that, when it comes to applications of other types, frames can be constructed in alternative ways and still satisfy definition 1. In particular, the definition leaves open which objects may play the role of possible data points. One could for instance take possible worlds in a metaphysically substantive sense or Carnapian state descriptions as elements of Ω. Which facts obtain in which worlds can be captured by Boolean variables. Any fact F can be represented by a Boolean variable $\tilde{F} : \Omega \to \{0, 1\}$ such that $\tilde{F}(\omega) = 1$ if and only if the fact F obtains in world ω. The distinction between dependent and independent variables then boils down to a distinction between fundamental and supervenient facts: independent variables represent fundamental facts that all other facts supervene on.[13] When constructing frames in such ways, the label 'possible data points' for the elements of Ω clearly must be taken with a grain of salt. The choice of this label is inspired by the central intended application. But in general one should think of possible data points simply as elementary possibilities at which sentences have truth values.

With frames defined, we can now bring language into the picture. We focus on propositional languages with countably many atomic sentences. We use 'p', 'q', 'r' (±indices) as meta-variables for atomic sentences. In practical applications, concrete sentences with internal structure such as 'age ≥ 20' or 'BMI $\in [18, 25]$' can be used as atomic sentences. We assume that the logical vocabulary includes the classical sentential connectives $\neg$, $\wedge$ and $\vee$. The material conditional ($\rightarrow$) and the biconditional ($\leftrightarrow$) can be included too, but we omit them for simplicity. Furthermore, the vocabulary contains symbols for would-counterfactuals ($\Box\!\!\rightarrow$) and might-counterfactuals ($\Diamond\!\!\rightarrow$). Formation rules for complex sentences are as usual. To distinguish between sentences with and without counterfactuals, we call sentences '*normal sentences*' iff they contain neither $\Box\!\!\rightarrow$ nor $\Diamond\!\!\rightarrow$. So when we

[13] In such a frame, facts may be stratified into many levels in the sense of List (2019) such that facts on any higher level supervene on facts of the next lower level. The hierarchy of levels does not have to be linear. There may be branches. All that variation semantics assumes is that there is a fundamental level.

simply say 'A is a sentence', A may contain occurrences of $\Box\!\!\rightarrow$ or $\Diamond\!\!\rightarrow$.

For sentences to have truth values, their language must be interpreted over a frame. This is what semantic models are for.

Definition 2. $(\Omega, \mathit{Var}, \mathit{Rel}, \mathit{Prop})$ is a **model** in variation semantics iff

1. (Ω, Var) is a frame;
2. *Rel* assigns a set of variables $\mathit{Rel}(p)$ to each atomic sentence p.
3. *Prop* assigns a set of possible data points $\mathit{Prop}(p)$, i.e. a proposition/intension, to each atomic sentence p.[14]

The set $\mathit{Prop}(p)$ is to be understood as the set of possible data points at which the atomic sentence p is true. For example, to the atomic sentence 'age $\geq$ 20' one would assign the set of all possible data points $\omega \in \Omega$ such that the variable *age* takes a value greater than or equal to 20 in ω.

The set $\mathit{Rel}(p)$ is to be understood as the set of variables relevant to the truth/falsity of p. Being relevant means being involved in making p true/false. For example, let p be the atomic sentence 'age $\geq$ 20'. If *Var* contains the variable *age*, we can let $\mathit{Rel}(p)$ be the set $\{\mathit{age}\}$, containing just this one variable. But there is some flexibility. If *Var* contains the variable *date of birth* as an independent variable in terms of which *age* is calculated, we can also set $\mathit{Rel}(p) = \{\mathit{age}, \mathit{date\ of\ birth}\}$. This alternative choice does not lead to any changes of truth values. It is okay but not necessary to close $\mathit{Rel}(p)$ under dependence.[15] If *Var* contains the variable *date of birth* but not *age*, then we have to set $\mathit{Rel}(p) = \{\mathit{date\ of\ birth}\}$ and specify $\mathit{Prop}(p)$ in terms of *date of birth*. In some cases, more than one variable is relevant to an atomic sentence. For example, the set of variables relevant to 'savings > amount requested' is $\{\mathit{savings}, \mathit{amount\ requested}\}$. In any case, the variables relevant to an atomic sentence can normally be read off the sentence itself when it comes to the central intended application of variation semantics.

So far, the key difference to similarity semantics is that models contain the two ingredients *Var* and *Rel* instead of a similarity relation. These ingredients are important to make the core idea of variation semantics precise. Recall that the

[14] Instead of *Prop* one could equivalently use a valuation assigning to each atomic sentence a truth value at each possible data point.

[15] The notion of dependence between variables is made precise in definition 4.

idea is to consider all variants of a given data point that make the antecedent true but otherwise coincide with the original data point.

One must be careful when fleshing out this idea. Let us return to the credit score example to illustrate a potential pitfall. Consider a data point ω_0 such that $age(\omega_0) = 60$, $income(\omega_0) = 38k$, $gender(\omega_0) = woman$, etc. Let A be the sentence 'age $= 40 \ \vee$ income $= 40k$'. Now what are the variants of ω_0 where A is true, other things being equal? A naive answer would be to say: exactly those possible data points at which A is true and that differ from ω_0 at most in the values of the variables *age* and *income* count as such variants. This captures all variants we want to consider. But it also includes unintended, problematic variants. For example, consider the data point ω^* where $age(\omega^*) = 40$ but $income(\omega^*) = 0$, while all other independent variables have the same values as in ω_0. This variant of the original data point satisfies the conditions in the naive answer. Nevertheless, this variant is not one we want to consider. We should not allow any further changes that are not involved in making the sentence A true. But ω^* contains an additional change that is not involved in making A true, namely the change of the value of the variable *income* to 0. It is the value of the variable *age* alone that makes A true in ω^*. If we require that the consequent of a counterfactual has to come out as true even under variants like ω^* that contain additional changes, it becomes much too hard for counterfactuals to be true. The naive account of A-variants is too wide. A more sophisticated approach is required.

Our approach uses the idea of exact truth-makers, which was introduced by Van Fraassen (1969) and has recently gained much attention in formal semantics.[16] To get an intuitive grasp of this idea, consider the example we just used to refute the naive account of A-variants. We saw that the value of the variable *income* is not involved in making the disjunction true in the data point ω^*. The value of *age* alone is what makes it true. More generally, what makes a sentence true at a given data point is typically not a full specification of values of all variables. Typically, a sentence is made true by what the values of some relevant variables are, so by partial data. For example, the atomic sentence 'income $\geq 40k$' is made true in some data point ω by the fact that $income(\omega) = 40k$, and it is made false in some other data point ω' by the fact that $income(\omega') = 20k$. For compound sentences, the situation is more complex. For example, the sentence 'age $= 40 \vee$ income $=$

[16] For an overview of truth-maker semantics see Fine (2017).

40*k*' can be made true (a) just by the fact that *age* takes the value 40, (b) just by the fact that *income* takes the value 40*k* or (c) by the fact that both *age* takes the value 40 and *income* takes the value 40*k*.

To make this talk of truth-making precise, we introduce the concept of a restriction of a data point ω to a set of variables V. We write '$\omega|_V$' for the restriction of ω to V. Intuitively, if ω is viewed as an assignment of values to all variables in *Var*, then $\omega|_V$ can be viewed as a partial assignment that agrees with ω except that it assigns values only to variables that are members of V. For our purposes, it suffices to define $\omega|_V$ simply as the ordered pair consisting of ω and V.[17] Using this concept, we can now define the relations '$\omega|_V$ is an **exact truth-maker** (short: **verifier**) of A in model $\mathcal{M}$' and '$\omega|_V$ is an **exact false-maker** (short: **falsifier**) of A in model $\mathcal{M}$' by simultaneous induction. These notions are defined along the lines of Van Fraassen (1969) and Fine (2012), except for the conditions regarding atomic sentences. Before turning to the definition, a small simplification is useful. Note that 'verifier' and 'falsifier' are model-relative concepts. But for the sake of brevity, we omit the phrase 'in model $\mathcal{M}$' from now on, wherever appropriate.

Definition 3. For all $\omega \in \Omega$, all $V \subseteq \mathit{Var}$, all atomic sentences p and all normal sentences A, B (recall that a *normal* sentence is one that does not contain a counterfactual conditional):

1. (a) $\omega|_V$ is a verifier of p iff $V = \mathit{Rel}(p)$ and $\omega \in \mathit{Prop}(p)$
   (b) $\omega|_V$ is a falsifier of p iff $V = \mathit{Rel}(p)$ and $\omega \notin \mathit{Prop}(p)$
2. (a) $\omega|_V$ is a verifier of $\neg A$ iff $\omega|_V$ is a falsifier of A.
   (b) $\omega|_V$ is a falsifier of $\neg A$ iff $\omega|_V$ is a verifier of A.
3. (a) $\omega|_V$ is a verifier of $A \wedge B$ iff there are $V_1, V_2 \subseteq \mathit{Var}$ such that $V = V_1 \cup V_2$ and $\omega|_{V_1}$ is a verifier of A and $\omega|_{V_2}$ is a verifier of B.
   (b) $\omega|_V$ is a falsifier of $A \wedge B$ iff $\omega|_V$ is a falsifier of A or $\omega|_V$ is a falsifier of B or $\omega|_V$ is a falsifier of $A \vee B$.
4. (a) $\omega|_V$ is a verifier of $A \vee B$ iff $\omega|_V$ is a verifier of A or $\omega|_V$ is a verifier of B or $\omega|_V$ is a verifier of $A \wedge B$.

[17] This is because data points are unstructured primitives in variation semantics. They are not defined at all. In particular, they are not defined as functions assigning values to variables. If they were, restrictions of data points could be defined as restrictions of functions.

(b) $\omega|_V$ is a falsifier of $A \vee B$ iff there are $V_1, V_2 \subseteq \mathit{Var}$ such that $V = V_1 \cup V_2$ and $\omega|_{V_1}$ is a falsifier of A and $\omega|_{V_2}$ is a falsifier of B.

The concept of a verifier is important for defining what an A-variant of a possible data point is. Those independent variables which are not involved in making a sentence A true in a given data point should keep their values in any A-variant of the data point. But this alone does not suffice. There can be cases where no independent variables are directly involved, as constituents of an exact truthmaker, in making A true. Yet, we cannot allow all independent variables to vary in such cases. To avoid too much variation, we need to keep the value of an independent variable X fixed if no dependent variable that is involved in making A true depends on X. So we need a notion of dependence to give a good definition of 'A-variant' avoiding further pitfalls.

Definition 4. Let X be an independent variable and Y a dependent variable. Then we say that Y **depends on** X iff there is a possible data point $\omega \in \Omega$ and a possible value $v \in \mathcal{V}_X$ of X such that Y takes different values in $\omega_{X:v}$ and ω. (Recall that $\omega_{X:v}$ is the possible data point at which X is set to the value v but all other independent variables take the same values as in ω.)[18]

Definition 5. Let $V \subseteq \mathit{Var}$. Then we say that B is **the basis of** V iff B is the set of all independent variables that are in V or that some variables in V depend on. We write '$\mathit{Basis}(V)$' for the basis of V.

With these definitions at hand, we can put forward our explication of 'A-variant'.

Definition 6. Let A be a normal sentence and let $\omega, \omega' \in \Omega$. Then we say that ω' is an ***A*-variant of** ω iff for some variable set $V \subseteq \mathit{Var}$:

1. $\omega'|_V$ is a verifier of A;
2. ω' agrees with ω regarding the values of all independent variables outside the basis of V, i.e. for all $X \in \mathit{Var}_i$: if $X \notin \mathit{Basis}(V)$, then $X(\omega') = X(\omega)$;
3. if A is true at ω, then ω' agrees with ω regarding the values of all variables in the basis of V.

[18] The existence and uniqueness of such a data point is guaranteed by our definition of 'frame' (condition 4).

There are two notable special cases: (a) If V does not contain any dependent variables, V is identical to its basis and the characterisation of A-variants can be simplified. The second condition becomes: ω' agrees with ω regarding the values of all independent variables that are not members of V. (b) If A is true at ω, then ω has only one trivial A-variant, namely itself. This is a consequence of the third condition, which corresponds to the centering assumption in similarity sphere semantics (Lewis, 1973a, section 1.3). So, the third condition captures the idea that if we seek to evaluate $A \mathrel{\Box\!\!\rightarrow} C$ at a given point ω but A is already true at that point, then there is no need to modify ω to make A true, and the truth value of $A \mathrel{\Box\!\!\rightarrow} C$ then depends only on the truth value of C at ω. But this special case rarely arises in practice because one typically uses a would-conditional only if one does not believe that its antecedent is satisfied in the actual circumstances.

With the idea of A-variants made precise, we now have a rigorous answer to the question under which conditions normal sentences and counterfactual conditionals are true at a given data point.

Definition 7. Let A and C be sentences, A a normal sentence, and let $\omega \in \Omega$.

1. A is true at ω iff for some variable set $V \subseteq \mathit{Var}$, $\omega|_V$ is a verifier of A.
2. $A \mathrel{\Box\!\!\rightarrow} C$ is true at ω iff C is true at all A-variants of ω.
3. $A \mathrel{\Diamond\!\!\rightarrow} C$ is true at ω iff C is true at some A-variant of ω.

Note that these truth conditions also cover nested counterfactuals. They cover counterfactuals with counterfactuals as *consequents*, e.g. $p \mathrel{\Box\!\!\rightarrow} (q \mathrel{\Box\!\!\rightarrow} r)$. However, as in any exact truth-maker semantics, our truth conditions do not cover counterfactuals that have counterfactuals as *antecedents*, e.g. $(p \mathrel{\Box\!\!\rightarrow} q) \mathrel{\Box\!\!\rightarrow} r$. This is so because we have defined A-variants only for normal sentences. Normal sentences have truth-makers in the technical sense of definition 3. But counterfactuals do not have truth-makers in this sense. They are not made true by what values particular variables take in a single data point. Instead, they are true (or false) in virtue of what holds in other possible data points that are suitably related to the actual data point. Nevertheless, we conjecture that the notion of an A-variant can be generalised to cover the case that A is a counterfactual. However, we do not see such a generalisation as a top priority because counterfactuals with counterfactuals as antecedents play virtually no role in practice. The other type of nested counterfactuals (those with counterfactuals as consequents) is more relevant in practice, and it can be handled by variation semantics already.

4 Advantages over alternatives

This section outlines what is novel about variation semantics. Its novel features are best seen by reviewing the advantages variation semantics has over its most prominent alternatives: similarity semantics and interventionist semantics.

4.1 Advantages vis-à-vis similarity semantics

Variation semantics has three advantages over similarity semantics: (a) It overcomes the problems of similarity semantics outlined in section 2. (b) It is simpler. (c) It yields more robust results.

The problems revisited. Let us first return to the examples regarding disjunctive antecedents, non-monotonic relationships and ranges to see how variation semantics overcomes the problems of similarity semantics.

Non-monotonic relationships. Recall the example from section 2.1: the sentence 'BMI $\in [t_1, t_2]$ $\Box\!\!\rightarrow$ risk not high' is true at a data point ω_0 according to similarity semantics although it should count as false. However, in a model of variation semantics that captures the scenario in question, the sentence can be shown to be false at the given data point ω_0. To see this, consider a data point where BMI takes a value between t_1 and t_1^* and where all other input variables have the same values as in ω_0. This then constitutes an A-variant of ω_0 (where A is 'BMI $\in [t_1, t_2]$') at which the risk score is still high and so the consequent is false. Hence, the counterfactual is false at ω_0.

Ranges. The general problem of ranges also disappears. In variation semantics, the counterfactual '(100 $\leq$ LDL (mg/dL) $\leq$ 200) $\Box\!\!\rightarrow$ LDL near optimal' is false at *any* possible data point in a model encoding the simple classification algorithm described in section 2.2. To see this, let ω be any possible data point. Let A be the antecedent of the conditional. Then there are always A-variants of ω where the consequent is false, no matter whether the LDL value lies within or outside the range specified in A. For example, one can consider the A-variant of ω where the LDL value is set to 200 mg/dL. Hence, the counterfactual is false at ω.

Disjunctive antecedents. Recall the scenario described in section 2.3. One can show that, in variation semantics, the sentence 'age = 40 $\vee$ income = $40k$) $\Box\!\!\rightarrow$ (credit score $\geq$ t)' is false at ω_0 in a semantic model capturing this scenario. This is so because ω_2 is an A-variant of ω_0, where A is the antecedent of the counterfactual in question. To see this, note that $\omega_2|_{\{age\}}$ is a verifier of A, and note that ω_2 agrees with ω_0 regarding the values of all input variables except *age*. However, the consequent is false in ω_2. So there is an A-variant of ω_0 where the consequent is false. Thus, the counterfactual is false at ω_0, as we argued it should be.

Simplicity. Variation semantics is in a sense simpler than similarity semantics because no similarity ordering is required. Instead of a similarity ordering on Ω, one has to specify a set of variables defined on Ω and also which variables are relevant to which atomic sentences. However, this is a straightforward task in the central intended applications of variation semantics. What the space of possible data points is and which variables to consider can normally be read off the specification of the machine learning model in question. Also, which variables are relevant to a given atomic sentence can normally be read off the sentence in question. In applications of similarity semantics, one has to specify a space of possibilities as well, but on top of that one must also choose a similarity ordering. And this choice is difficult. There is in general no unique answer to what the right similarity ordering is in machine learning (cf. Molnar, 2019, section 9.3). So there is a sense in which similarity semantics has an additional free parameter compared to variation semantics.

Robustness. As a consequence of its greater simplicity, variation semantics yields more robust results than similarity semantics. In similarity semantics, the truth values of counterfactuals crucially depend on the choice of a similarity ordering, e.g. on how variables are weighted when making overall comparisons of possible data points. If one changes the similarity ordering, this can change the truth values of counterfactuals. In contrast, variation semantics yields the same results independently of any preferred similarity ordering. In the central intended applications of variation semantics, the constituents of a semantic model are determined by the algorithm and the sentences under consideration. Hence, the truth values of counterfactuals are determined by the parameters of the application alone. They do not depend on difficult decisions about the right standard of similarity.

4.2 Advantages vis-à-vis interventionist semantics

In variation semantics, one evaluates counterfactuals relative to models that specify functional relationships between dependent and independent variables. In this regard, it is similar to interventionist semantics, which offers truth conditions for counterfactuals in terms of causal models in the structural equations framework. This similarity raises the question how variation semantics compares to interventionist semantics.

Interventionist semantics has been developed on the basis of Judea Pearl's work on causal models (Galles & Pearl, 1998; Halpern, 2000; Pearl, 2009), and it has been brought in its most general form by Briggs (2012). Causal models are popular not only in philosophy but also in computer science, for example as tools for testing whether algorithms satisfy fairness criteria such as counterfactual fairness (Kusner et al., 2017).

In interventionist semantics, one evaluates a counterfactual conditional of the form '$X = a \mathrel{\Box\kern-1.5pt\rightarrow} Y = b$' relative to a causal model by adding the equation '$X = a$' to the causal model while removing any other equation from the model that determines the value of X, to avoid inconsistency. This step is referred to as a 'surgical intervention'. If '$Y = b$' holds in the new model resulting from the intervention, then '$X = a \mathrel{\Box\kern-1.5pt\rightarrow} Y = b$' is true in the original model. Otherwise it is false.

Variation semantics and interventionist semantics both explicate the idea of making the antecedent true by setting the values of certain variables while keeping other things equal.[19] But they differ in how 'other things' is understood. We believe that interventionist semantics is not in general adequate for counterfactual reasoning about algorithmic decisions for the following reason: when dependent variables are relevant to the antecedent of a counterfactual, surgical interventions change the underlying causal model (algorithm) to make the antecedent true. A surgical intervention sets the values of some variables by force and severs the links between these variables and any variables they depend on by removing equations from the model. Surgical interventions are useful for some purposes. But they are a problem when evaluating counterfactuals regarding the behaviour of algorithms

[19] We conjecture that they yield the same truth values for equational counterfactuals with antecedents that concern only independent variables (e.g. '$X = a \mathrel{\Box\kern-1.5pt\rightarrow} Y = b$', where X is an independent variable).

because the algorithm must remain unchanged for a good counterfactual explanation of an algorithm's output. One seeks to find out how possible alternative inputs would affect the algorithm's outputs rather than what would happen if one changed dependencies within the algorithm itself.

To illustrate this, consider a model where the output variable depends on the one hand directly on the independent variables but, on the other hand, also on an intermediate interaction variable. For example, consider a very simple model that predicts the expected level of economic marginalisation of a person in a particular society (Y) in terms of three independent variables

(X_1) gender (0: man, 1: woman),

(X_2) race (0: white, 1: black),

(X_3) disability status (0: able-bodied, 1: disabled)

and an interaction variable capturing multiple discrimination (D).[20]

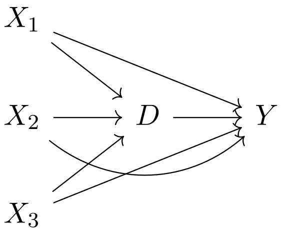

Figure 2: mathematical model with intermediate interaction variable.

The model is given by two equations:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D$$

$$D = \alpha_1 X_1 X_2 + \alpha_2 X_1 X_3 + \alpha_3 X_2 X_3$$

Note that D takes the value 0 unless at least two of the three independent variables take the value 1. So, D takes the value 0 for anybody who is not subject to multiple discrimination. For simplicity of illustration, we assume that the parameters $\alpha_1, \ldots, \alpha_3$ and $\beta_1, \ldots, \beta_4$ are all equal to 1. Given this simplification,

[20] This example is inspired by the work of Bright et al. (2016) and Kim et al. (2019). That only the variables *gender*, *race* and *disability status* are used in this model and that they are treated as binary merely serves to simplify the exposition.

D takes the value 1 for people who are subject to double discrimination (black women, women with disability, black people with disability); and it takes the value 3 for people who are subject to triple discrimination (black women with disability). As a consequence, the output variable Y takes the value 3 in cases of double discrimination and the value 6 in the case of triple discrimination.

Now suppose a white man without disability makes the following claim regarding this model:

(C5) If I were subject to double discrimination, my expected level of economic marginalisation would be 1. ($D = 1 \,\Box\!\!\rightarrow Y = 1$)

According to interventionist semantics, this counterfactual is true. We will now explain why this is so and why this result poses a problem for interventionist semantics.

First, to capture the scenario in question, we set $X_1 = 0, \ldots, X_3 = 0$ ("white man without disability"). Given these values of the independent variables, the second model equation yields $D = 0$. So we get $Y = 0$, using the first equation. This is what the model yields for the actual data point. Now, to evaluate the counterfactual C5 at this data point according to interventionist semantics, one performs a surgical intervention in the model. One sets $D = 1$ but leaves $X_1, \ldots, X_3$ all set to 0. To avoid inconsistency, one removes the original equation for D from the model. So one severs the links between the independent variables and the variable D, which thereby ceases to be an interaction variable in the modified model (figure 3).

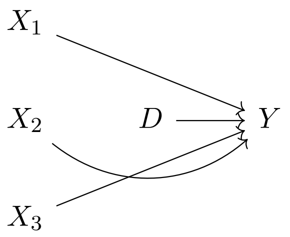

Figure 3: mathematical model after intervention.

Then one calculates the value of Y on this basis. The result is $Y = 1$. So, the intervention that makes the antecedent ('$D = 1$') true, results in the truth of the consequent ('$Y = 1$'). Hence, C5 is true according to interventionist semantics.

This result is problematic. In effect, we have calculated the expected level of economic marginalisation for a white man without disability who is artificially counted as being subject to double discrimination without satisfying the necessary conditions for that. Counterfactual explanations of algorithmic decisions should not rely on such reasoning. Note that the original model would not yield $Y = 1$ for any possible input where $D = 1$, i.e. for any person who is actually subject to double discrimination. The reason is that, for D to take the value 1, exactly two of the independent variables $X_1, \ldots, X_3$ must also take the value 1, by the second equation of the model, which defines D. But in that case, the first equation yields $Y = 3$. This suggests that '$D = 1 \boxright Y = 1$' should count as false and '$D = 1 \boxright Y = 3$' should come out as true instead if one seeks to describe the behaviour of the model as it is.

Variation semantics captures this intuition. In variation semantics, the underlying model remains unchanged. One considers all $(D = 1)$-variants of the given data point, and one calculates the value of Y for all these possibilities. In every such variant, exactly two of the input variables $X_1, \ldots, X_3$ take the value 1. As a consequence, the sentence '$D = 1 \boxright Y = 3$' comes out as true in variation semantics and '$D = 1 \boxright Y = 1$' as false.

5 Logical consequence and hyperintensionality

We have now seen how variation semantics differs from similarity semantics and interventionist semantics regarding truth values. But it also differs with regard to the relation of logical consequence (entailment). The definition is as usual.

Definition 8. Let A be any sentence and T any set of sentences. Then we say that T **entails** A in variation semantics iff for every model $\mathcal{M}$ and every possible data point ω in $\mathcal{M}$: if all sentences in T are true at ω, then A is also true at ω.

One difference to similarity semantics concerns counterfactuals with disjunctive antecedents. As the example in section 2.3 shows, $(A \vee B) \boxright C$ does not entail $A \boxright C$ in similarity semantics. But the situation is different in variation semantics.

Proposition 1. *$(A \vee B) \boxright C$ entails both $A \boxright C$ and $B \boxright C$ in variation semantics.*

Proof. Without loss of generality, we consider only the case of $A \mathrel{\Box\!\!\rightarrow} C$. To see that this sentence is entailed by $(A \vee B) \mathrel{\Box\!\!\rightarrow} C$, simply apply the definition of entailment and the fact that every A-variant of a possible data point is also an $(A \vee B)$-variant of it. This fact follows from the definition of 'A-variant' along with condition 4.(a) of the definition of 'verifier'. □

Variation semantics also differs from interventionist semantics in terms of logical consequence. In interventionist semantics, Modus Ponens is not valid for counterfactuals. The reason is that an intervention regarding a dependent variable alters the underlying model even if the variable is set to a value it had anyway (Briggs, 2012). In contrast:

Proposition 2. *Modus Ponens is valid for counterfactuals in variation semantics. That is:* $\{A,\ A \mathrel{\Box\!\!\rightarrow} C\}$ *entails* C *in variation semantics.*

Proof. Use the simple fact that ω is an A-variant of itself if A is true at ω. □

Moreover, variation semantics captures important invalidities that are commonly taken to distinguish counterfactuals from strict conditionals. For example:

Proposition 3. *Antecedent Strengthening is not valid in variation semantics. That is:* $A \mathrel{\Box\!\!\rightarrow} C$ *does not entail* $(A \wedge B) \mathrel{\Box\!\!\rightarrow} C$ *in variation semantics for pairwise distinct atomic sentences* A, B, C.

Proof. Take a simple linear model with three variables (X_1, X_2, Y) such that $Y(\omega) = X_1(\omega) + X_2(\omega)$ for all $\omega \in \Omega$. Consider the data point ω_0 such that $X_1(\omega_0) = 0 = X_2(\omega_0)$. Let *Prop* and *Rel* be such that

sentence	true at	relevant variables
A	those $\omega \in \Omega$ with $X_1(\omega) = 1$	X_1
B	those $\omega \in \Omega$ with $X_2(\omega) = 1$	X_2
C	those $\omega \in \Omega$ with $Y(\omega) = 1$	Y

Note that $A \mathrel{\Box\!\!\rightarrow} C$ is true at ω_0. Setting X_1 to 1, leaving other things equal, results in Y taking the value 1. However, $(A \wedge B) \mathrel{\Box\!\!\rightarrow} C$ is false at ω_0. Setting both X_1 and X_2 to 1, other things being equal, results in Y taking the value 2. □

Another feature of variation semantics that is important from a logical point of view is that it is a hyperintensional system. More specifically:

Proposition 4. *Not all sentences that are tautologically equivalent can be substituted in a counterfactual without changing the truth value of the counterfactual in variation semantics.*

Proof. Consider the model and the data point ω_0 from the proof of Proposition 3. The sentence '$X_1 = 1 \mathbin{\Box\!\!\rightarrow} Y = 1$' is true at ω_0 in this model. We are using the intended interpretation of the atomic sentences in this example. So, for instance, *Prop*('$X_1 = 1$') $= \{\omega \in \Omega : X_1(\omega) = 1\}$, and *Rel*('$X_1 = 1$') $= \{X_1\}$. Note that '$X_1 = 1$' is tautologically equivalent to '$(X_1 = 1 \vee (X_1 = 1 \wedge X_2 = 10))$'.[21] However, '$(X_1 = 1 \vee (X_1 = 1 \wedge X_2 = 10)) \mathbin{\Box\!\!\rightarrow} Y = 1$' is false at ω_0, according to variation semantics. To see this, suppose for reductio that it is true. Then '$(X_1 = 1 \wedge X_2 = 1) \mathbin{\Box\!\!\rightarrow} Y = 1$' must also be true by Proposition 1. However, this is not the case, as explained in the proof of Proposition 3.

□

To sum up, the consequence relation of variation semantics differs from the consequence relations of similarity semantics, interventionist semantics and any semantics that treats counterfactuals as strict conditionals.

6 Range of applicability & generalisations

We now address a potential concern about variation semantics, namely that it has a limited range of applicability. The worry is that variation semantics allows us to evaluate only counterfactuals about the behaviour of algorithms and other mathematical models and that, therefore, it cannot be used to evaluate counterfactuals of other types such as counterfactuals about natural or social phenomena (e.g. 'If Nixon had pressed the button, there would have been a nuclear war between the USA and the USSR'). This raises the question whether there is a generalisation of variation semantics that can deal with counterfactuals of any type and, if so, whether variation semantics is made obsolete by the more general semantics.

Our response is two-fold: (a) There are in fact several ways of generalising variation semantics. However, in many cases, there are still good reasons for preferring variation semantics when it comes to concrete applications. (b) The range

[21] Fine (2017) uses an example of the same structure to illustrate the problem of disjunctive antecedents.

of applicability of variation semantics is wider than it might seem at first glance, and it is not even clear that its generalisations have a significantly greater range of applicability. Let us elaborate on both points.

Generalisations. There are at least two semantics that can be viewed as generalisations of variation semantics. Priest's (2008, chapter 5) general semantics for conditionals is one. It is a possible world semantics. A model is given by a set of worlds together with a family of accessibility relations, one relation R_A for each sentence A. A conditional $A \mathrel{\Box\!\!\rightarrow} C$ is true at a world w_0 iff C is true at every R_A-accessible world (i.e. at every world w_1 with $w_0 R_A w_1$). The suggested reading of '$w_0 R_A w_1$' is 'A is true at w_1, which is, ceteris paribus, the same as w_0' (Priest, 2008, paragraph 5.3.3). However, the intuitive idea behind this reading is left almost entirely unexplicated in Priest's framework.[22] Only very weak conditions are imposed on R_A.[23] Therefore, Priest's framework is enormously general. It encompasses not only variation semantics but also similarity semantics and even a semantics of counterfactuals as strict conditionals as special cases. Strict conditional semantics is obtained by taking all the A-worlds as R_A-accessible worlds. Similarity semantics is obtained by taking the *closest* A-worlds as the R_A-accessible worlds. Variation semantics is obtained by identifying possible data points with worlds and taking A-variants as the R_A-accessible data points.

Fine's (2012) abstract truth-maker semantics is another framework that can be seen as a generalisation of variation semantics. It is a hyperintensional semantics based on possible states (rather than possible worlds) and on parthood relations between states. States serve as verifiers and falsifiers of sentences. To give truth conditions for counterfactuals, Fine works with transition relations $s_0 \rightarrow_w s_1$, where s_0, s_1 are states and w is a world. The intuitive reading of '$s_0 \rightarrow_w s_1$' is 'state s_1 is a possible outcome of imposing the change s_0 on the world w'. A conditional $A \mathrel{\Box\!\!\rightarrow} C$ is classified as true at a world w iff for any states s_0 and s_1 such that s_0 is an exact truth-maker of A and $s_0 \rightarrow_w s_1$, s_1 contains an exact truth-maker of C. Parthood relations and transition relations are taken as primitive in Fine's framework, and only weak conditions are imposed on them to guarantee maximum generality. One can construct models of this framework

[22] One can view the definition of 'A-variant' as a substantive explication of this idea.

[23] The conditions are: (1) A is true at w_1 if $w_0 R_A w_1$. (2) $w R_A w$ if A is true at w. (Priest, 2008, section 5.5)

from models of variation semantics. To do so, one can take partial data as states, i.e. possible assignments of values to (not necessarily all) variables, and define a parthood and transition relation between them in terms of notions of variation semantics.[24]

The extreme generality of the frameworks of Priest and Fine has advantages but also disadvantages. On the one hand, these frameworks are very flexible and can be adapted in numerous ways because they place only very weak constraints on models. On the other hand, generality comes at a cost. The general frameworks can be difficult to apply when it comes to determining the truth values of counterfactuals in concrete cases. Doing so requires the choice of a semantic model. But due to the highly abstract form of models in each framework, most of the difficult work lies in constructing a suitable model.

When it comes to concrete applications, variation semantics may be preferable to its generalisations for at least two reasons. First, it is continuous with modelling practices in the sciences and in engineering as far as its conceptual architecture is concerned, and therefore it is easier to construct suitable semantic models in it. Second, its range of applicability is much wider than it may seem. Let us elaborate on each point.

Continuity with scientific practice. Models specifying functional relationships between dependent and independent variables are the bread and butter of natural and social scientists as well as engineers in many fields, including policy-relevant areas such as economics and epidemiology. Variation semantics allows us to evaluate counterfactuals based on such models. It uses possible data points and a common mathematical conception of variables. Thus, variation semantics is much closer to the practice of scientific modelling than similarity semantics, Fine's truth-maker semantics and other highly abstract semantics popular among philosophers. It constitutes a user-friendly foundation for counterfactual reasoning based on mathematical models that is applicable in many fields.

Applicability. It is not true that variation semantics is applicable only to counterfactuals about the behaviour of mathematical models. It can also be applied

[24] There is a natural parthood relation on variable assignments, e.g. $(X_1 = a_1)$ is a part of $(X_1 = a_1, X_2 = a_2)$. A transition relation can be defined along the lines of the definition of 'A-variant': $s_0 \rightarrow_w s_1$ iff s_1 is a data point that contains partial data s_0 as a part and agrees with the data point w regarding the values of all independent variables outside the basis of the set of variables occurring in s_0.

to counterfactuals about natural or social phenomena. And it does not appear to be less suitable for this purpose than the two more general frameworks. To show this, let us return to the example from the beginning of this section:

(C6) If Nixon had pressed the button, there would have been a nuclear war between the USA and the USSR.

Since this sentence is about historical scenarios and not about the behaviour of a particular mathematical model, one might think that it cannot be handled by variation semantics. But this is not correct.

One can evaluate C6 by constructing a model that represents dependencies between possible historical situations. Let us construct a very simple model using only Boolean variables (taking only the values 0 or 1) for illustration. Let N and W be Boolean variables such that '$N = 1$' represents that Nixon presses the button, and '$W = 1$' represents that there is a nuclear war between the USA and the USSR. Then '$N = 0$' and '$W = 0$' represent the negations of these statements. Furthermore, let L, B_1 and B_2 be Boolean variables such that '$L = 1$' represents that the USA launch nuclear missiles on the USSR; '$B_1 = 1$' represents that particular background conditions under which the USA launch nuclear missiles obtain; and '$B_2 = 1$' represents that particular background conditions under which the USSR launch nuclear missiles obtain. Suppose these variables are related as follows, for all $\omega \in \Omega$:

$$L(\omega) = 1 \text{ iff both } N(\omega) = 1 \text{ and } B_1(\omega) = 1.$$

$$W(\omega) = 1 \text{ iff both } L(\omega) = 1 \text{ and } B_2(\omega) = 1.$$

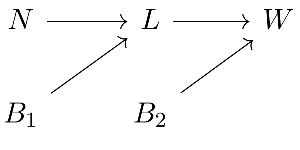

Figure 4: a simple historical model.

Based on this model, one can evaluate C6 (in symbols: $N = 1 \mathrel{\Box\!\!\rightarrow} W = 1$). Consider the possible data point ω such that $N(\omega) = 0$, $B_1(\omega) = 1$ and $B_2(\omega) = 1$, assuming that this assignment of values to the independent variables in question is historically accurate. There is only one ($N = 1$)-variant of ω, namely the data point ω' such that $N(\omega') = 1$, $B_1(\omega') = 1$ and $B_2(\omega') = 1$. Note that '$W = 1$' is true at ω' by the conditions above. So, C6 is true at ω.

Admittedly, this is an extremely simplistic model. It merely serves to illustrate the general idea. But one could construct more complex and realistic models. Constructing a good semantic model is essential because the truth value of the counterfactual we seek to evaluate depends on the chosen model. However, how a good model should look like in this case is not a logical question but rather a question for historians and social scientists.

But note that more general semantics do not fare better when it comes to evaluating counterfactuals such as C6. To evaluate a sentence, one needs a semantic model. And we do not see good reasons to think that constructing a good semantic model is easier in much more abstract frameworks than variation semantics. Consider Fine's truth-maker semantics. It is at least as difficult to see what a set of states, a parthood relation and a transition relation on states should be like to serve as a good basis for evaluating the counterfactual C6 as it is to see which historical variables might play a role and what their dependencies might be. So, when it comes to determining truth values of concrete counterfactuals, we doubt that more general frameworks such as Fine's and Priest's possess significant advantages over variation semantics.

7 Conclusions and outlook

Our main conclusions are: (a) Variation semantics constitutes a good framework for evaluating counterfactual conditionals in explanations of algorithmic decisions. It it more suitable for this purpose than similarity semantics, interventionist semantics and the frameworks of Fine and Priest. (b) Since its range of applicability extends far beyond counterfactuals about the behaviour of algorithms, variation semantics is a genuine alternative to similarity semantics and interventionist semantics.

7.1 Implications for practice and computer science

Variation semantics has potential for practical applications in all areas concerned with counterfactuals about the behaviour of algorithms, including machine learning and law. When it comes to practical applications, a computational implementation of variation semantics would be helpful. This raises the question whether one can write a computer programme, based on variation semantics, that can calculate the truth value of a given counterfactual about the behaviour of a given algorithm. Such a programme promises to be of practical use for individuals or organisations who seek to evaluate counterfactual conditionals to understand or contest algorithmic decisions.

7.2 Implications for philosophy and logic

There are implications for the debate about the right methodology for assessing causal effects of social categories such as race and gender and the closely related issue of how counterfactuals about a person's gender or race can be evaluated. The widely used interventionist approach has recently come under increasing criticism in this debate (Hu, 2022; Kasirzadeh & Smart, 2021).[25] This paper supports critics of interventionism. Interventionist semantics presupposes that any variable is manipulable in isolation. In a surgical intervention, one may for instance change the value of the variable *race* from 'black' to 'white' or vice versa without changing the values of any other variables, even if *race* depends on other variables in the model in question.[26] We have argued in section 4.2 that conclusions drawn from such interventions can be problematic. Moreover, we have provided an alternative to interventionist semantics. If *race* is a dependent variable in a model, then — according to variation semantics — it cannot be changed without also changing some variables it depends on. Although variation semantics is not intended as an approach to causal inference, it might be a useful starting point for critics of interventionism to develop rigorous frameworks capturing their views.

From a logical point of view, the novel conceptual architecture of variation semantics may be of interest for two reasons.[27] (a) It can be used as a starting point

[25] See Malinsky and Bright (2021) for an argument in favour of the manipulability of race.

[26] For example, the variable *race* might depend on "ancestry, self-awareness of ancestry, public awareness of ancestry, culture, experience of privilege or oppression, and subjective self-identification" (Malinsky & Bright, 2021).

[27] In term of the structure of models and its conceptual construction, variation semantics differs

for developing new hyperintensional semantics. (b) It challenges a widespread view about the design of hyperintensional semantics.

Regarding the development of new semantics, a first natural step is to extend the ideas presented in this paper to first-order languages. To do so, one could replace possible data points with possible worlds, where each possible world comes with an associated domain of objects; and one could replace variables with attributes, where an n-ary attribute could be viewed as a function that assigns to each world an n-ary relation in the mathematical sense[28] over the domain of that world. Moreover, one could take dependence relations between attributes as primitive rather than defining them as in the basic version of variation semantics presented here. Various types of dependence relations could be considered, including supervenience, grounding or causal dependence relations.

Regarding the design of hyperintensional systems more generally, one of the most prominent approaches is to start with possible states (or situations) rather than possible worlds.[29] Possible worlds can then be defined as particular maximally-consistent states. Fine (2012) claims that much hinges on adopting a state-based approach rather than a possible worlds semantics, for example to overcome the problem of disjunctive antecedents. But this seems to be an illusion. Variation semantics belongs to the family of possible worlds semantics, but it can solve Fine's problems too. This suggests that a different choice of primitives is equally viable: possible data points (or possible worlds), variables (or attributes) and an assignment of relevant variables (or attributes) to atomic sentences. Using this or a similar architecture to develop hyperintensional semantics for logical operators other than counterfactual conditionals could be a fruitful strand of research in philosophical logic and AI.

from traditional semantics for hyperintensional operators such as truth-maker semantics (Fine, 2017), situation semantics (Barwise & Perry, 1983), impossible worlds semantics (Nolan, 1997) and various semantics for relevance logics (e.g. Anderson & Belnap, 1975; Anderson et al., 1992; Routley & Meyer, 1973).

[28] By a 'relation in the mathematical sense' we mean a set of ordered n-tuples.

[29] Barwise and Perry (1983) were among the first to endorse such an approach. A state-based semantics for counterfactuals was first conceived and developed by Fine (1975, 2012). But also other important notions have been explicated in this tradition, for example 'permission' (Anglberger & Korbmacher, 2020) or 'synonymy' (Hornischer, 2020). For excellent overviews, see Fine (2017) and Leitgeb (2019, section 1).

7.3 Outlook

Desiderata for further research include: (a) developing a deductive system for variation semantics and proving its soundness and completeness; (b) studying the formal relationships between variation semantics and other semantics in more detail; and (c) extending variation semantics to give an account of probabilistic counterfactual reasoning. As for the latter point, a natural step to extend models is to endow Ω with a probability measure. Then one can assign (global) probabilities to counterfactuals since counterfactuals express propositions.[30] But perhaps more interestingly, one can define (local) counterfactual probabilities. If a would-counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is false at a point ω while the corresponding might-counterfactual $A \mathrel{\Diamond\!\!\rightarrow} C$ is true — so if C is true at some but not all A-variants of ω —, then it is natural to ask how *probable* C is within the set of A-variants of ω. This suggests defining, at any point ω, a local counterfactual probability of C being the case if A were the case (notation: '$P_\omega(C\|A)$') by setting $P_\omega(C\|A) := P(\mathit{Prop}(C)|\langle A\rangle_\omega)$, where P is the background probability measure on Ω, $\mathit{Prop}(C)$ is the set of points at which C is true and $\langle A\rangle_\omega$ is the set of A-variants of ω. This definition of local counterfactual probabilities is novel in several respects. It constitutes an alternative to Balke & Pearl's explication of counterfactual probabilities in terms of causal models (1994a, 1994b). It also differs from suppositional accounts of probabilities of counterfactuals such as Bradley's (2021) and from Leitgeb's (2012a, 2012b) probabilistic semantics for counterfactuals. It seems that variation semantics offers a fresh starting point for thinking about counterfactual probabilities and probabilities of counterfactuals.

[30] The global probability of a counterfactual is the probability of the set of data points in Ω at which it is true. Pearl (cf. 2009, chapter 7) defines probabilities of counterfactuals essentially like that as well.

References

Anderson, A. R., & Belnap, N. D. (1975). *Entailment: The Logic of Relevance and Necessity* (Vol. Volume I). Princeton University Press.

Anderson, A. R., Belnap, N. D., & Dunn, J. M. (1992). *Entailment: The Logic of Relevance and Necessity* (Vol. Volume II). Princeton University Press.

Anglberger, A., & Korbmacher, J. (2020). Truthmakers and Normative Conflicts. *Studia Logica*, *108*, 49–83. https://doi.org/10.1007/s11225-019-09862-5

Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations - a survey. *ArXiv*, *abs/1911.07749*.

Balke, A., & Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds and applications. *Uncertainty proceedings 1994* (pp. 46–54). Elsevier.

Balke, A., & Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, 230–237.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. MIT Press.

Bradley, R. (2021). Probabilities of counterfactuals. *Argumenta*, *12*, 179–193. https://doi.org/10.14275/2465-2334/202112.bra

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*, 139–166. https://doi.org/10.1007/s11098-012-9908-5

Bright, L. K., Malinsky, D., & Thompson, M. (2016). Causally interpreting intersectionality theory. *Philosophy of Science*, *83*(1), 60–81. https://doi.org/10.1086/684173

Clift, A. K., Coupland, C. A. C., Keogh, R. H., Diaz-Ordaz, K., Williamson, E., Harrison, E. M., Hayward, A., Hemingway, H., Horby, P., Mehta, N., Benger, J., Khunti, K., Spiegelhalter, D., Sheikh, A., Valabhji, J., Lyons, R. A., Robson, J., Semple, M. G., Kee, F., . . . Hippisley-Cox, J. (2020). Living risk prediction algorithm (qcovid) for risk of hospital admission and mortality from coronavirus 19 in adults: National derivation and validation cohort study. *BMJ*, *371*. https://doi.org/10.1136/bmj.m3731

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, & H. Trautmann (Eds.), *Parallel problem solving from nature – ppsn xvi* (pp. 448–469). Springer International Publishing.

Fine, K. (1975). Critical notice on david lewis' counterfactuals. *Mind*, *84*(335), 451–458. http://www.jstor.org/stable/2253565

Fine, K. (2012). Counterfactuals without possible worlds. *Journal of Philosophy*, *109*(3), 221–246. https://doi.org/10.5840/jphil201210938

Fine, K. (2017). Truthmaker semantics. In B. Hale, C. Wright, & A. Miller (Eds.), *A Companion to the Philosophy of Language* (2nd edition).

Galles, D., & Pearl, J. (1998). An Axiomatic Characterization of Causal Counterfactuals. *Foundations of Science*, *3*, 151–182. https://doi.org/10.1023/A:1009602825894

Geddes, L. (2021). Scientists question NHS algorithm as young people called in for jab. *The Guardian*. https://www.theguardian.com/society/2021/mar/09/scientists-question-nhs-algorithm-as-young-people-called-in-for-jab

Halpern, J. Y. (2000). Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research*, *12*, 317–337. https://doi.org/10.1023/A:1009602825894

Hornischer, L. (2020). Logics of synonymy. *Journal of Philosophical Logic*, *49*(4), 767–805. https://doi.org/10.1007/s10992-019-09537-5

Hu, L. (2022). Interventionism in Theory and in Practice in the Social World [Manuscript].

Kaminski, M. (2018). The right to explanation, explained. *Berkeley Technology Law Journal*, *34*, 189.

Kasirzadeh, A., & Smart, A. (2021). The Use and Misuse of Counterfactuals in Ethical Machine Learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 228–236. https://doi.org/10.1145/3442188.3445886

Kim, E. J., Parish, S. L., & Skinner, T. (2019). The impact of gender and disability on the economic well-being of disabled women in the united kingdom: A longitudinal study between 2009 and 2014. *Social Policy & Administration*, *53*(7), 1064–1080. https://doi.org/https://doi.org/10.1111/spol.12486

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Leitgeb, H. (2012a). A probabilistic semantics for counterfactuals. part a. *The Review of Symbolic Logic*, *5*, 26–84. https://doi.org/10.1017/S1755020311000153

Leitgeb, H. (2012b). A probabilistic semantics for counterfactuals. part b. *The Review of Symbolic Logic*, *5*, 85–121. https://doi.org/10.1017/S1755020311000165

Leitgeb, H. (2019). HYPE: A System of Hyperintensional Logic (with an Application to Semantic Paradoxes). *Journal of Philosophical Logic*, *48*, 305–405. https://doi.org/10.1007/s10992-018-9467-0

Lewis, D. (1973a). *Counterfactuals*. Harvard University Press.

Lewis, D. (1973b). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, *2*(4), 418–446. https://doi.org/10.1007/BF00262950

List, C. (2019). Levels: Descriptive, explanatory, and ontological. *Noûs*, *53*(4), 852–883. https://doi.org/https://doi.org/10.1111/nous.12241

Malinsky, D., & Bright, L. K. (2021). On the causal effects of race and mechanisms of racism [Manuscript].

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Mothilal, R. K., Sharma, A., & Tan, C. (2019). Explaining machine learning classifiers through diverse counterfactual explanations. *CoRR*, *abs/1905.07697*. http://arxiv.org/abs/1905.07697

Nolan, D. (1997). Impossible Worlds: A Modest Approach. *Notre Dame Journal of Formal Logic*, *38*(4), 535–572. https://doi.org/10.1305/ndjfl/1039540769

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Priest, G. (2008). *An introduction to non-classical logic: From if to is* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511801174

Routley, R., & Meyer, R. K. (1973). The Semantics of Entailment I. In H. Leblanc (Ed.), *Truth, syntax and semantics* (pp. 194–243). North-Holland.

Stalnaker, R. C. (1968). A Theory of Conditionals. In N. Rescher (Ed.), *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)* (pp. 98–112). Oxford: Blackwell.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001.

Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 465–474. https://doi.org/10.1145/3097983.3098039

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. https://doi.org/10.1145/3287560.3287566

Van Fraassen, B. C. (1969). Facts and Tautological Entailments. *Journal of Philosophy*, *66*(15), 477–487. https://doi.org/10.2307/2024563

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review.

Verma, S., Dickerson, J., & Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/jopp.12262

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887.

Wexler, R. (2017). When a Computer Program Keeps You in Jail: How Computers are Harming Criminal Justice. *New York Times*. https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html