Causality, Counterfactuals, and Belief.
More Means-End Philosophy

Franz Huber

Under contract with Oxford University Press.

Contents

Preface

This book is the second of two volumes on belief and counterfactuals. It consists of five of a total of eleven chapters. ... ... Finally, while merely a change in terminology, I should perhaps note that, throughout the second volume, I follow my own suggestion from the first volume of referring to subjective probabilities not anymore as what they are not, viz., degrees of belief, but as what they are: degrees of certainty.

I am very grateful to Holger Andreas, Sander Beckers, Mario Günther, Alan Hájek, Joe Halpern, Christopher Hitchcock, Alice Huang, Joshua Knobe, Thomas Kroedel, David Papineau, Brian Skyrms, Wolfgang Spohn, Joost Vennekens for comments on an earlier version of this book.

Toronto, July 2023

Parts of chapter 7 rely on and, with permission of Springer, as well as the editors and Mentis, respectively, reuse material from Huber (2014; 2017), as well as Huber (2016), respectively.

Parts of chapter 8 rely on and, with permission of the co-author and Wiley, *The British Journal for the Philosophy of Science*, *Philosophy of Science*, Mentis and the editors, as well as Springer, respectively, reuse material from Kroedel & Huber (2013), Huber (2011), Huber (2012), Kroedel & Huber (2013), Huber (2016), and Huber (2017), respectively.

Parts of chapter 9 rely on and, with permission of Cambridge University Press, reuse material from Huber (2013).

Parts of chapter 10 rely on and, with permission of Springer, reuse material from Huber (2015).

Chapter 7

Counterfactuals

In this chapter I will first discuss the use of intuitions in contemporary analytic philosophy. Then I will outline an approach to philosophy that does not overly rely on intuitions: means-end philosophy. Next I will present my view of possible worlds: modal idealism. Finally, I will state the royal rule – a normative principle relating descriptive normality and conditional belief. In the presence of the thesis that beliefs ought to obey the ranking calculus, as well as the truth conditions of default conditionals and counterfactuals in terms of descriptive normality, this principle determines the logical postulates satisfied by default conditionals and counterfactuals. This chapter heavily relies on Huber (2014; 2016; 2017).

7.1 Intuitions

There are at least two distinct uses of intuitions in philosophy. Often one describes a particular case and then form intuitions about this case. Such intuitions about a case are judgments about what is true in the case described. On the basis of these intuitions the person who has them can evaluate a general principle that the particular case instantiates, or fails to instantiate. One can, of course, also evaluate claims other than general principles that (one takes to) relate to the particular case, just as one can form intuitions not only about particular cases, but arbitrary matters. However, for the purposes of this chapter we can focus on intuitions about particular cases which instantiate, or fail to instantiate, general principles. Gettier (1963) uses intuitions in this way when he describes the particular case of Smith, and then uses his intuition about this case to conclude that the general principle that justified true belief is sufficient for knowledge is false.

Gettier (1963) shares a judgment, but does not offer an argument (other than that the general principle is false if it fails to be instantiated in the particular case). To the extent that Gettier (1963)'s readers share his judgment, they too can use their judgments to conclude that justified true belief is insufficient for knowledge. If so, Gettier (193)'s readers, at the time or their reading, use the word 'know' in a way similar to the way Gettier used it when writing his article. Alternatively, the concept each reader, at the time of their reading, has of the word 'know' is similar to Gettier's concept of it when writing his article.

Stalnaker (1994) uses intuitions in a seemingly similar way. He describes the particular case of an agent who initially believes Bizet and Satie are French and Verdi is Italian. Then he considers what this agent would believe if she received the information ϕ that Bizet and Verdi are compatriots, as well as what she would believe if she received the information $\phi \wedge \xi$ that Bizet and Verdi, as well as Satie and Verdi are compatriots. Stalnaker (1994: 19) uses his intuitions about this particular case to conclude "that it would be unreasonable to require rational belief revision to conform to [the general principle of rational monotonicity]." In the present context, the latter principle says the following: if an agent would believe ψ, but not $\neg\xi$ if she came to believe the information ϕ (but no logically stronger information), then she should believe ψ if she came to believe the information $\phi \wedge \xi$ (but no logically stronger information). Here, ψ is the proposition that Satie is French. (As an aside, note that Stalnaker 1994: 19 considers what the agent would believe if she came to *learn* rather than believe these propositions. Presumably, learning is different from believing, but for the sake of argument we grant that this does not make a difference.)

Like Gettier (1963), Stalnaker (1994) shares a judgment, but does not offer an argument. However, fom the means-end perspective adopted in this book and the first volume, there is a difference between these two uses of intuitions. Gettier (1963) is concerned with a general principle about what knowledge *is*, whereas Stalnaker (1994) is concerned with a general principle about how beliefs *should be* revised. This means Stalnaker (1994) and his readers may use their judgments to conclude one of two things: either the agent in question acts irrationally by failing to revise her beliefs in the way she should given her ends; or she acts rationally by revising her beliefs in the way she should given her ends, but these ends are not the ones the normative principle of rational monotonicity is hypothesized upon and a means to attaining. What neither Stalnaker (1994) nor his readers may conclude is that the normative principle of rational monotonicity is false. This is so because its truth consists in the obtaining of an established means-end relation that is not affected by particular cases or anyone's intuitions about them (see chapter 5).

Belief does, but knowledge does not, figure in (the main argument place of) norms because belief is, but knowledge is not, an action. For this reason one might think that describing particular cases and sharing one's intuitions about them may be the only way to do philosophy when one is studying knowledge – or rather, uses of the word 'know' on particular occasions. In what follows I will attempt to show, first, that if this were the case, this would be bad news, as different philosophers have intuitions that are too different or too unspecific; and, second, that this is not the case. My topic is not knowledge, though, but the modality expressed by counterfactuals: counterfactuality.

Philosophers often rely on intuitions when theorizing about counterfactuals. However, intuitions regarding counterfactuals are notoriously shaky. To illustrate, consider the debate between Stalnaker (1968; 1981) and Lewis (1973a) about the validity of the so-called law of conditional excluded middle: for any two sentences α and γ, if it were the case that α, it would be the case that γ, or, if it were the case that α, it would not be the case that γ. In symbols:

0. $\vdash (\alpha \mathrel{\Box\!\!\rightarrow} \gamma) \vee (\alpha \mathrel{\Box\!\!\rightarrow} \neg\gamma)$

According to Stalnaker (1968) this principle is logically valid. Lewis (1973a: 77ff) disagrees and brings the following alleged counterexample:

C It is not the case that if Bizet and Verdi were compatriots, Bizet would be Italian; and it is not the case that if Bizet and Verdi were compatriots, Bizet would not be Italian; nevertheless, if Bizet and Verdi were compatriots, Bizet either would or would not be Italian.

Stalnaker (1981: 91ff) defends an analysis which says that both of the following two counterfactuals (from Quine 1950: §3)

C1 If Bizet and Verdi had been compatriots, Bizet would have been Italian.

C2 If Bizet and Verdi had been compatriots, Verdi would have been French.

"are indeterminate – neither true nor false. It seems to me that the latter conclusion is clearly the more natural one. I think most speakers would be as hesitant to deny as to affirm either of the conditionals, and it seems as clear that one cannot deny them both as it is that one cannot affirm them both. Lewis seems to agree that unreflective linguistic intuition favors this conclusion." (Stalnaker 1981: 92) The reason for Stalnaker's last claim is that Lewis (1973a: 80) writes: "I want to say [C], and think it probably true [...]. But offhand, I must admit, it does sound like a contradiction. Stalnaker's theory does, and mine does not, respect the opinion of any ordinary language speaker who cares to insist that it is a contradiction."

As Stalnaker (1981: 92) points out, "it would be arbitrary to require a choice of one of [C1 and C2] over the other, but [...] this is not at issue. What is at issue is what conclusion about the truth values of the counterfactuals should be drawn from the fact that such a choice would be arbitrary." The conclusion drawn by Lewis (1973a) is that both C1 and C2 are false and, hence, that conditional excluded middle is not logically valid. The conclusion drawn by Stalnaker (1968; 1981) is that both C1 and C2 are indeterminate and that conditional excluded middle remains without counterexample and is logically valid. End of discussion.

An example from the more recent literature is the discussion between Gillies (2007) and Moss (2012) about where to draw the line between the semantics and the pragmatics of counterfactuals. Here is the relevant background. Lewis (1973a: 10), referring to Sobel (1970), uses so-called Sobel sequences to argue against the analysis of the counterfactual as a strict conditional. Sobel sequences are examples such as:

S1 If Sophie had gone to the parade, she would have seen Pedro.

S2 If Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro.

Lewis (1973a) assumes that counterfactuals such as these can be jointly true. This is not so if the counterfactual is a *strict* conditional, i.e., a necessary material conditional. Lewis (1973a) concludes that the counterfactual is a *variably* strict conditional, i.e., a strict conditional whose strictness varies with the antecedent.

Gillies (2007) considers *reverse* Sobel sequences that von Fintel (2001: 130) attributes to Irene Heim:

S2 If Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro.

S1 If Sophie had gone to the parade, she would have seen Pedro.

For Gillies (2007: 332) "this sounds for all the world like a contradiction." Gillies (2007) then goes on to argue that the counterfactual is a strict conditional, but one whose truth values interact with context in such a way that the order in which two counterfactuals are asserted matters.

Moss (2012) defends the analysis of the counterfactual as a variably strict conditional. She admits that reverse Sobel sequences are "generally infelicitous." However, unlike Gillies (2007), she does not consider them to be contradictory. Moss (2012) first considers non-conditional sentences and notes that in some cases the order in which they are uttered matters:

M1 That animal was born with stripes.

M2 But cleverly disguised mules are not born with stripes.

For her this conversation is felicitous, while the reversed one is not:

M2' Cleverly disguised mules are not born with stripes.

M1' But that animal was born with stripes.

Moss (2012: 568) claims that "[o]ur intuitions about [M'] point towards a general principle governing assertability [... which ...] tells us that if a speaker cannot rule out a possibility made salient by some utterance, then it is irresponsible of her to assert a proposition incompatible with this possibility."

As before, "[w]hat is at issue is what conclusions about the truth values of the counterfactuals should be drawn" from the fact that reverse Sobel sequences are considered to be contradictory and infelicitous, respectively. The conclusion drawn by Gillies (2007) is that the truth values of counterfactuals depend on the order in which they are uttered. The conclusion drawn by Moss (2012) is that the assertability conditions, but not the truth values, of counterfactuals depend on the order in which they are uttered. End of discussion.

Both examples[1] illustrate a common pattern, although what it is depends on one's view. One view is that two philosophers have different intuitions about a particular case: Lewis (1973a) intuits falsity of C1 and C2, while Stalnaker (1981) intuits arbitrariness of choosing between C1 and C2; Gillies (2007) intuits contradictoriness of S2-S1, while Moss (2012) intuits infelicity of S2-S1. Another view is that two philosophers share an intuition – the arbitrariness of choosing between C1 and C2; the infelicity of S2-S1 – but disagree on its details: Lewis (1973a) intuits falsity in addition to arbitrariness, but Stalnaker (1981) does not (and, perhaps, intuits indeterminacy in addition to arbitrariness); Gillies (2007) intuits contradictoriness in addition to infelicity, but Moss (2012) does not (and, perhaps, intuits unassertability in addition to infelicity). A third view is that two philosophers share an intuition – the arbitrariness of choosing between C1 and C2; the infelicity of S2-S1 – but use it to evaluate different general principles: the validity of conditional excluded middle versus the indeterminacy of the truth of counterfactuals; a semantic principle that governs truth value assignments versus a pragmatic principle that governs assertability.

[1] Another example is the discussion between Lewis (1973a; 1981) and Stalnaker (1968; 1981) versus Kratzer (1981) and Pollock (1976) about the semantic principle of Comparability. The latter says, roughly, that any two worlds can be compared with respect to their similarity to the actual world. Cf. Lewis (1981: sct. 5).

These are bad news, no matter one's view. Allegedly, our intuitions are the "evidence" that decides between rival philosophical theories (Pust 2000). On the first and second view, we do not agree what the evidence is, let alone what it says. On the third view, we agree what the evidence is, but even if we agree what it says, it vastly underdetermines philosophical theory. Our intuitions are too different or they are too unspecific. Either way, they do not decide between rival philosophical theories.

Besides discussions in philosophical logic and the philosophy of language this affects other discussions involving counterfactuals: in epistemology kowledge is analyzed in terms of counterfactuals (Nozick 1981, Roush 2005); in metaphysics it is causation (Collins et al. 2004, Paul & Hall 2013); in the general philosophy of science it is dispositions (Mumford 1998); in the philosophy of biology there are special "biologically normal" counterfactuals (Weber 2021); outside philosophy, psychologists study regret and responsibility with counterfactuals (Connolly et al. 1997), while historians use counterfactuals in thought experiments (Reiss 2009).

One reaction is to go experimental (Knobe & Nichols 2008) and see which intuitions are more widespread. On my view this does not help much because information about how intuitions are distributed across various populations does not settle philosophical issues. Indeed, on my view this makes things worse, as it makes us focus on what is *not* the arbiter in philosophical debates: intuition. Instead, this arbiter is argumentation from premises, and with assumptions, both of which need to be stated clearly so that everyone can judge them for themselves, whether by intuition or otherwise. As shown in the first volume, philosophy can be done other than by describing particular cases and sharing one's intuitions about them when the question is what should be done. As shown in the next section, philosophy can also be done in this way when the question is what is the case.

7.2 Means-end philosophy

In the way I will continue to engage with it, epistemology is a normative discipline that studies how agents should believe on the hypothesis that they have certain ends such as holding true and informative beliefs. In the way I will engage with it, metaphysics is subordinate to epistemology insofar as metaphysical theses are necessary conditions for the satisfiability of epistemological norms. Given the instrumentalist understanding of normativity, this means that metaphysical theses are necessary conditions for the possibility of attaining ends from epistemology. How exactly is this "transcendental" metaphysics supposed to work?

My topic is the modality expressed by counterfactuals: counterfactuality. The metaphysical theses I am interested in include, among other things, the logical postulates satisfied by counterfactuals and default conditionals. More specifically, they concern the properties of the semantic ingredient of the truth conditions of counterfactuals and default conditionals.

In order to show that these metaphysical theses are necessary conditions for the possibility of attaining a cognitive end, I will formulate a normative principle on conditional beliefs that pertains to this semantic ingredient and is subject to the hypothesis that the agent has said cognitive end. The normative principle itself needs to be justified by being shown to actually be a means to attaining said cognitive end. This will be attempted in section 8.1. In addition, there are assumptions that I need to state clearly so that you can judge them for yourself. Whether I find these assumptions intuitive does not matter, as my argument is not intended to be one by authority. Nor does it matter for the validity of my argument whether you have the cognitive end in question. The reason is that the normative principle is a hypothetical, not a categorical imperative. Let me state my assumptions, then, before turning to the normative principle.

My first assumption is that counterfactuals and default conditionals express propositions that are true or false. For dissenting, expressivist views regarding counterfactuals see Edgington (2008) and Spohn (2013; 2015). The latter has it that counterfactuals express propositions relative to the agent's conditional beliefs and a partition.

My second assumption is that counterfactuals and default conditionals express propositions that are true or false at possible worlds. For a dissenting, possible states view regarding counterfactuals see Fine (2012a).

My third assumption is that the truth conditions of counterfactuals and default conditionals at possible worlds are as follows. Let α and γ be sentences from a formal language $\mathcal{L}$ that are evaluated for truth at the elements of a non-empty set of possible worlds W. A default conditional $\alpha \Rightarrow \gamma$ is true at w in W if, and only if, γ is true at all worlds v in W (i) at which α is true and (ii) which are most normal from the point of view of w. A counterfactual $\alpha \boxrightarrow \gamma$ is true at w if, and only if, (1) γ is true at all worlds at which α is true and which are most normal from the point of view of w; and – if w is itself less normal from its point of view than the most normal (from the point of view of w) worlds at which α is true – (2) γ is true at all worlds at which α is true and which are at least as normal (from the point of view of w) as w. Thus, the counterfactual $\alpha \boxrightarrow \gamma$ says that, normally – and even if things are not normal, as long as they are not less normal than the way things actually are – if α is true, then so is γ.

The world of evaluation w may, but need not be among the worlds which are most normal from its point of view. If it is, the counterfactual $\alpha \mathrel{\square\!\!\rightarrow} \gamma$ is true at w if, and only if, the default conditional $\alpha \Rightarrow \gamma$ is. Otherwise the counterfactual may be false while the default conditional is true. However, the converse case cannot occur.

There are several counterfactuals that may contradict each other, as we will see in section 8.3. These are causal or interventionist counterfactuals, backtracking counterfactuals, and spurious or acausal counterfactuals. However, these different counterfactuals can be given a unified semantic treatment by assuming that their differences lie not in their truth conditions, but in what is held fixed, in addition to the antecedent or *if*-part α, in evaluating normality at a possible world: the causal structure plus what is causally upstream of α, the causal structure alone, and nothing that is causal (in the narrow sense of effective causation), respectively.

The central semantic ingredient of the truth conditions just stated is normality at a possible world. In the presence of these truth conditions, its properties across all possible worlds determine the logical postulates satisfied by counterfactuals and default conditionals. However, the role played by these truth conditions must not be underestimated. Different truth conditions and a semantic ingredient with different properties can determine the same logical postulates for counterfactuals, as we will see in section 7.4. For this reason it is worth noting that one can argue for these truth conditions as follows.

In order to give a possible worlds semantics for necessity and possibility claims, counterfactuals, normality claims, and default conditionals, we need some semantic ingredient in addition to the non-empty set of possible worlds and the evaluation function that evaluates atomic sentences for truth at possible worlds. Arguably, we need the very semantic ingredient of normality at a possible world. The reason is that the one alternative semantic ingredient we need on independent grounds – viz., (objective, single-case) probability at a possible world – does not provide an adequate semantics for default conditionals (see Halpern 2003/2017: ch. 8, especially pp. 295ff).

Now enter Occam's razor, which urges us to avoid multiplying entities without necessity. In particular, it requires us to not additionally postulate the existence of a second semantic ingredient – unless we have to. We do not have to, though. Default conditionals and counterfactuals can be given a possible worlds semantics in terms of normality at a world, and so can non-conditional necessity, possibility, and normality claims. To complete the argument, note that Occam's razor is a hypothetical imperative that is justified by being shown to be a means to attaining the truth efficiently (Kelly 2007 and Kelly et al. 2016).

Normality, or typicality, is understood in a purely descriptive, not evaluative or prescriptive sense (Bear & Knobe 2017) and a singular, not generic or statistical sense (Sytsma et al. 2012). To illustrate, consider the following example (from Reiter 1980): normally, if Tweety is a bird, it can fly. This means Tweety can fly in the most normal worlds in which it is a bird. Identifying the normality of a proposition with the normality of the most normal possible worlds comprising it, this means it is more normal for Tweety to be a bird and be able to fly than for it to be a bird and not be able to fly. Sometimes default conditionals are inferred from *generic default rules* such as that birds can normally fly, but penguins cannot. In this case the default conditional that, normally, Tweety can fly if it is a bird can be inferred only if one has no information about Tweety that contradicts the claim that it can fly. For instance, one cannot infer that it is more normal for Tweety-the-penguin to be a bird and be able to fly than for it to be a bird and not be able to fly. Similarly, the default rule that U.S. presidents normally do not tweet does not allow one to conclude that it is more normal for Donald Trump to be U.S. president and not tweet than it is for him to be U.S. president and tweet – just as the generic information that, statistically speaking, U.S. presidents are likely male does not allow one to conclude that the first female U.S. president is likely male. Generic default rules and statistical information are formulated in terms of generic variables that are defined on a population of individuals. In contrast to this, singular default conditionals and claims about single-case probabilities are formulated in terms of singular variables (Huber 2018: sct. 10.7). The question under what conditions the former license inferences to the latter is a variant of the reference class problem (Huber 2018: sct.s 10.2, 10.8). We will study in chapter 10 under which conditions the truth values of default conditionals and counterfactuals can be reliably inferred from statistical information.

With these assumptions being stated, let us turn to the normative principle, the royal rule, which relates normality to conditional belief. The general idea behind the royal rule is that, absent further information, alethic modality constrains or guides doxastic modality. An approximation of it in terms of default conditionals says that an agent should believe a proposition C on the assumption that she is certain of the proposition A, as well as the default conditional that, normally, if A, then C, but no overriding information. The idea is that, absent overriding information, default conditionals constrain or guide conditional beliefs. More precisely, the royal rule says that one ought to disbelieve a particular proposition to a particular grade on the assumption that it is, in a purely descriptive sense, abnormal to this grade for this proposition to be true, but no information that is not entirely about normality.

The royal rule is similar in spirit to Lewis' (1980) "principal principle" which relates another alethic modality, viz. chance, to another doxastic modality, viz. conditional degree of certainty. This principle says that an agent's initial degree of certainty in a proposition C ought to equal x given that the chance equals x that C is true and, perhaps, further "admissible" information, but no inadmissible information. With the help of a couple of assumptions about what information is admissible (historical information and information on how chance depends on history), the principal principle entails that chances behave how an agent's initial conditional degrees of certainty ought to behave. Now, initial degrees of certainty – and, given the ratio formula, initial conditional degrees of certainty – ought to obey the probability calculus. Therefore, chances do so as well. So, probabilism, i.e., the thesis that degrees of certainty ought to obey the probability calculus, including the ratio formula, and the principal principle have a consequence that is about chance – namely that chances are probabilities. While, presumably, this claim is also in agreement with the subjective intuitions of many, there is no need to appeal to the latter in order to defend this claim. Given a couple of assumptions, probabilism and the principal principle do this on their own. This illustrates how two normative principles from epistemology can entail a metaphysical thesis.

Suppose we can also justify these two normative principles by showing them to be means to attaining ends one may have. Probabilism can perhaps be justified by the Dutch Book argument (Ramsey 1926, de Finetti 1937).[2] The principal principle can perhaps be justified in some other way (Pettigrew 2013). If so, the thesis that chances are probabilities is a consequence of probabilism and the principal principle which in turn can be justified by being shown to be means to attaining ends one may have. Intuitions are certainly useful as a heuristics in arriving at these normative principles, as well as the auxiliary assumptions, and in considering various metaphysical theses. However, one need not appeal to intuitions in order to defend the metaphysical thesis.

If all this works out, the metaphysical thesis that chances are probabilities is a necessary condition for the possibility of attaining certain ends one may have. Given that one has these ends, one ought to satisfy those norms. Yet one can satisfy those norms only if things are a certain way. So, means-end philosophy tells one what metaphysical theses one is committed to by pursuing various ends.

[2]Joyce (1998; 2009)'s gradational accuracy dominance argument does not quite suffice for present purposes for two reasons. First, it does not cover the ratio formula. This is not fixed easily because conditional accuracy is as unintelligible as conditional truth. In contrast to this, conditional bets are as intelligible as bets. Second, Joyce (2009: 279) appeals to the principal principle in defense of his assumptions about inaccuracy.

In the same transcendental way I want to derive the properties of descriptive normality across all possible worlds from the royal rule and the normative thesis that beliefs ought to obey the ranking calculus, including the difference formula. The latter normative thesis is justified by the consistency argument from chapter 5. A justification of the royal rule will be attempted in section 8.1. Given the truth conditions assumed of default conditionals and counterfactuals, the properties of normality across all possible worlds determine the logical postulates satisfied by default conditionals and counterfactuals. These logical postulates are expected to approximate the logical postulates philosophers have proposed on the basis of their subjective intuitions. However, we do not have to rely on those intuitions in order to justify these postulates. Instead, we obtain them as consequences of two normative principles from epistemology, as well as assumptions about the truth conditions of default conditionals and counterfactuals.

While irrelevant to the validity of this means-end argument, to be convinced by it and, hence, accept its conclusion, one needs to accept the assumptions made, as well as pursue the ends the normative principles are hypothesized upon and means to attaining. Those who do not are given information about means-end relationships for which they may have little or no use.

7.3 Modal idealism

A factual, or non-modal, language, or system of representation, allows one to say *that* something is the case: it is raining; the streets are wet. A modal language presupposes a factual language and allows one to say, in addition, *how* something factual is the case: possibly it is raining; typically (normally) it is not; if it was, the streets would be wet. Of course, what is a modal claim in one language may be a factual claim in another. For each factual language there is exactly one linguistic, conceptual, or representational entity that accurately and maximally specifically – that is, as completely as the factual language allows – describes, or represents, reality: the *actual factual world* for the factual language under consideration.

The actual factual world is not real, as modal realism would have it (Lewis 1986a). Rather, it is a mind-dependent construct that is somewhat similar to a state description (Carnap 1947a); it is an *idea*. Apart from the actual factual world there are many merely possible factual worlds. These are all the descriptions in the factual language under consideration that maximally specifically, but *in*accurately describe reality. Factual worlds give rise to factual propositions which we can formally represent as sets of factual worlds. So much for a factual language.

In a modal language we can say more than in a factual language. In particular, this is true for an alethically modal language (for the time being we restrict the discussion to languages with one modality). In such a language, we can say not only that it is not raining, but also that it could have been raining, that this would have been atypical, and that, if it had been, the streets would have been wet.

Formally, alethically modal propositions can be represented as sets of modal worlds. The latter consist of a factual component and a modal component. In our case of descriptive normality the modal component of a modal world specifies what is normal at its factual component. The modal component does *not* specify what is normal in reality. Instead, it specifies what is normal at some factual world, which may or may not be the actual factual world. Since factual worlds are not real, but ideas, alethically modal propositions are not about reality, but about ideas.

For each alethically modal language and each factual world there is exactly one modal component that accurately and maximally specifically describes how – in the sense of this modality – things are the case at this factual world. In our case of descriptive normality this means that for each factual world there is exactly one modal component that accurately and maximally specifically describes what is normal to what degree at this factual world. This modal component is determined not by reality, but by the factual world and the alethically modal language under consideration. Reality has a say in this only insofar as it has a say in which factual world is actual, and which language one is speaking.

Factual and modal worlds are relative to a language in the broadest sense of this term. They are not real in any way that is independent of language, thought, conceptualization, and representation. We humans find ourselves representing, conceptualizing, thinking about, and talking about reality in terms of what is and is not, what could have been, and what would have been. Yet these *not*s, *could*s, and *would*s have no reality themselves. Rather, they belong to the language we use to describe reality and, thus, to the realm of the mental. As its name suggests, only reality is real. It, however, may not have the conceptual structure that our conceptualizations of it (e.g., Tarski 1935's "Folge von Gegenständen") have.

Suppose it is neither raining nor snowing, that it could have been raining and snowing, that this would have been atypical, and that, if it had been, the streets would have been wet. On the present view there is nothing real that is described by, corresponds to, or makes true these *not*s, *or*s, *and*s, *could*s, *atypical*s, and *would*s. Rather, these words and their meanings are confined to our representation of reality in thought and language (needless to say, we cannot think, let alone talk about reality without some representation).

Just as thinking and talking about reality are dependent on a language, so is truth. This is why these thoughts and claims can have truth values without there being anything real that is described by them, corresponds to them, or makes them true. What is and is not true, what could have been true, and what would have been true depend on reality because it depends on reality which factual world is actual. Yet what is true *also* depends on the language these propositions are expressed in and dependent on. Alethically modal claims can express truths – and not merely beliefs or other cognitive states – without there being any mind-independent modal reality described by them, corresponding to them, or making them true. The reason is that they can be understood as claims about ideas.

Idealism about alethic modality is a position between modal realism and modal expressivism in the tradition of Hume (1739; 1748) (see Price 2008). Like the modal expressivist, the modal idealist does not locate the modalities in reality, but in the mind. Like the modal realist, the modal idealist does not interpret the modalities as expressing cognitive states, but as propositions that are true or false. Of course, the nature of propositions differs on the realist and idealist accounts.

Alethic modalities are ideas. However, alethic modalities are not ideas with mind-independent reality. Instead, they are *our* ideas. We humans find ourselves conceptualizing reality in terms of *could*s, *atypical*s, and *would*s, just as we find ourselves conceptualizing reality in terms of *not*s, *or*s, and *and*s. Different beings who also think (in the broadest sense of this term) about reality, as well as humans in different times and places, may find themselves conceptualizing or representing reality in different terms or ideas – say, because they have different abilities and limitations, or because they have different ends. There is no right or wrong here, nor any necessity. All there is is a more or less useful for various purposes.

Modal idealism stands in the combinatorial tradition of Wittgenstein (1921) (see also Skyrms 1981) that Lewis (1986a) terms "linguistic ersatzism." Among others, this approach faces the problem of descriptive power (see also Sider 2002): distinct possible worlds are not distinguished (in addition, some possible worlds are omitted entirely). In its most pressing form this problem results from the assumption that there could have been properties other than the ones there actually are. Since these possible properties are not actually instantiated, we cannot name them directly. At best we can describe them indirectly by specifying the roles they play. However, in this case we cannot distinguish possible worlds in which distinct properties swap roles: worlds that differ only in regard to which of two or more distinct, merely possible properties play a specified role. (The problem also arises for individuals instead of properties, but in this form it is less pressing, as only haecceitists will be troubled by it.)

More generally, the combinatorialist needs to specify the building blocks of her actual world, as well as the rules for recombining these to form new, merely possible worlds. The challenge is to find a way to make it true that there could have been building blocks other than the ones there actually are.

The modal idealist addresses this challenge by denying the presumption that there is one privileged set of building blocks plus rules of recombination. Instead, there are other languages besides the one she finds herself speaking. For her to say that there could have been building blocks other than the ones there actually are is to say that there are languages other than the one she finds herself speaking that have different building blocks (and, perhaps, different rules of recombination). Specifically, for her to say that there could have been properties (and individuals) other than the ones there actually are is to say that there are languages which name or characterize uniquely properties (and individuals) that have no name or unique description in her language. I subscribe to these claims. In fact, if, as I believe, there are languages that do no conceptualize reality in terms of individuals and properties, but carve reality differently or do not *carve* it at all, then there are mere possibilia that are neither individuals nor properties – possibilia that Lewis (1986a) omits. If it is true that there could have been building blocks other than properties and individuals, Lewis (1986a)'s modal realism gets the facts of modality wrong.

The combinatorialist needs to specify the building blocks of her actual world, as well as the rules for recombining these to form new, merely possible worlds. In the case of linguistic ersatzism the latter specify in a syntactic, proof-theoretic or semantic, model-theoretic manner which recombinations of the linguistic building blocks are consistent. Together with the notion of maximality (that is specified set-theoretically), consistency determines possibility. Lewis (1986a) claims that linguistic ersatzism needs alethically modal primitive vocabulary. The reason is that, allegedly, consistency gets the case of contingent laws of nature wrong, no matter whether it is logical consistency or consistency with additional postulates: if no additional postulates are added, "contingent laws" are not laws; if additional postulates are added, "contingent laws" are not contingent.

We will not have any use for laws of nature, contingent or otherwise, but only (in chapter 9) for necessary default conditionals – unless these are what laws of nature are (but see Woodward 2003: ch. 5). If one has use for them, though, the modal idealist can accommodate by pointing out that the additional postulates added to one language need not be added to every language – and cannot be, if there are languages that do not conceptualize reality in terms of individuals and their properties and relations.

Finally, while truth of a proposition in a possible world can be identified with set-theoretic elementhood of the latter in the former, I have not said anything yet about how the actual factual world is selected among all possible factual worlds, nor how the actual modal world is selected among all possible modal worlds. One option is to take this relation between possible worlds, reality, and language, as well as the talk of '(in)accuracy' from the beginning of this section, as primitive. However, for an instrumentalist a more promising alternative is to identify the actual world as the possible world that provides the most useful description for the purposes of the language under consideration (i.e., the purposes of the speakers of this language *qua* speakers of this language). This allows us to understand language in the broadest sense of this term that goes beyond the languages studied in logic and linguistics and includes those that do not conceptualize reality in terms of individuals and properties. (I assume that there is a uniquely most useful world for every language merely for the sake of simplicity.)

What can be said in favor of modal idealism? We humans have ideas, as not only realists and expressivists admit, or so I will assume, but also philosophers who deny the existence of reality. Now enter Occam's razor, which urges us, in giving truth conditions for alethically modal claims, to avoid postulating the existence of another kind of entity – unless we have to. We do not have to, though.

This appeal to Occam's razor leaves unspecified the particular form that modal idealism takes. Specifically, it leaves unspecified its "nested" character. The latter is applicable also to realist positions, as well as indeterministic alethic modality, and has it that modal claims qualify factual claims. To bring out a feature of this nested character, let me first state in more detail Lewis (1980)'s principal principal. Then I will discuss a problem that arises from it in connection with the metaphysical thesis of Humean supervenience – a problem that does not arise if the nested character of modality is acknowledged. This discussion will serve also as a foil for the next section.

Principal Principle (Lewis 1980). *An ideal cognitive agent's initial degree of certainty function* Pr *should be a regular probability measure satisfying the ratio formula such that for all times t, all (standard or non-standard) real numbers x, all propositions A, and all propositions $E_{A,t}$ that are admissible for A at t:*

$$\Pr(A \mid ch_t(A) = x \cap E_{A,t}) = x$$

Here, $ch_t(A) = x$ is the proposition that the chance at t that A is true exists and equals x.

The idea is that, for propositions A for which chances $ch_t(A)$ exist at various times t, their chances at these times guide or constrain the ideal cognitive agent's initial, or a priori, conditional degrees of certainty in these propositions for all (and only those?) conjunctive conditions $ch_t(A) = x \cap E_{A,t}$ such that the conjunct $E_{A,t}$ is admissible for these propositions at these times.

For instance, if you are not certain of anything other than the tautology, then your initial conditional degree of certainty that a coin will land on heads given that this coin is fair and will be tossed until it lands on heads or tails should equal a half. Of course, your initial conditional degree of certainty that the coin will land on heads given that it is fair and will land on tails should be zero. So, the latter conditional degree of certainty should *not* be equal to the chance that the coin lands on heads according to the hypothesis that it is fair. This is so because the information that the coin will land on tails "overrides" the information that the coin is fair for the question whether the coin lands on heads.

Information that is admissible for proposition A at time t is information that affects the ideal cognitive agent's initial degree of certainty in A *only*, if at all, by affecting her initial degree of certainty in what the chance of A at t is or if this chance exists. On the conditional theory of conditional degree of certainty (section 5.2), this notion of conditional doxastic (in-) dependence is to be understood, in exact parallel to section 6.3, as counterfactual (in-) dependence of the relevant doxastic states. A different way of putting this is to say that the information $ch_t(A) = x$ "screens off" any information that is admissible for A at t. It does so in the same way as the information that Biden has won Georgia in the 2020 U.S. presidential election screens off the information that Trump is leading the polls in Georgia on the eve of election day for my degrees of certainty in who will win the 2020 U.S. presidential election. The latter information about Trump's lead, while, in general, relevant to my degrees of certainty in who will win the 2020 U.S. presidential election, does not affect these degrees of certainty anymore in the presence of the former information about Biden's win.

What information is admissible for which propositions at which times depends on one's view of indeterministic alethic modality. Lewis (1980) assumes historical information about times no later than t to be admissible at t for all propositions. In particular, the complete history of any possible world w up to t, $H_{w,t}$, is admissible at t for all propositions. In addition, Lewis (1980) assumes information about how chances at any time t depend on complete histories up to t to be admissible at all times for all propositions. In particular, any possible world w's theory of chance, T_w, is admissible for all propositions at all times. Finally, admissibility at any time (for any proposition) is assumed to be closed under Boolean combinations.

Lewis (1980) identifies the theory of chance of a possible world w with the conjunction, or intersection, of all "history-to-chance conditionals" of the form 'if H_t, then $ch_t(A) = x$' that are true at w. Here, H_t specifies a possible complete history up to time t. The conditional operator in question can be Lewis (1973a)'s counterfactual. However, as long as this operator satisfies modus ponens, this is not required for the argument to follow.

In the presence of these assumptions the principal principle implies that an ideal cognitive agent's initial conditional degree of certainty function Pr should be such that for all possible worlds w, all times t, and all propositions A for which the chance at w at t exists:

$$ch_{w,t}(A) = \Pr(A \mid H_{w,t} \cap T_w)$$

This consequence says that the chance distribution of any possible world w at any time t, $ch_{w,t}$, equals an ideal cognitive agent's initial conditional degree of certainty function whose condition is the conjunction of the complete history of w up to t and w's theory of chance. It implies the metaphysical thesis that chances obey the probability calculus because probabilism, including the ratio formula, says that the initial conditional degrees of certainty an ideal cognitive agent should have do so. This means we do not have to postulate this metaphysical thesis, but can derive it from two normative principles, as well as assumptions about admissibility. A means-end argument for this metaphysical thesis is completed, if the two normative principles can be justified by being shown to be means to attaining ends one may have. That chances are probabilities then has been shown to be a necessary condition for the possibility of attaining ends one may have, provided the assumptions about admissibility hold.

A mathematical oddity is that Lewis (1980) is working with non-standard probability measures to ensure that all propositions $ch_t(A) = x$ receive non-zero probability in the sense of Pr (Bernstein & Wattenberg 1969). Otherwise the conditional probabilities $\Pr(A|ch_t(A) = x \cap E_{A,t})$ may not be defined. In order to avoid this problem, Spohn (2010) considers propositions of the form $x < ch_t(A) < y$, an idea due to Skyrms (1980). Another option may be to work with Popper-Rényi functions (section 4.3). Such manoeuvres will not be required in the case of the royal rule. A philosophical oddity arises from Lewis (1986b)'s Humean supervenience assumption. According to it, chances supervene on local matters of particular fact. This claim has no motivation in modal idealism (or modal expressivism). Presumably, it is meant to ease the ontological burden of modal realism. The nested character of modality helps to get clear about what this claim amounts to, as well as why it fails.

Let F be the set of all factual worlds, where a factual world f now completely specifies a possible totality of local matters of particular fact. A modal world w is a pair (f_w, ch_w) consisting of a purely factual component f_w from the set of factual worlds F and a purely (indeterministic alethic) modal component ch_w from the set of possible chance measures CH for F that are defined on some algebra $\mathcal{A}_F$ over F. The set of all modal worlds W is (a subset of) the Cartesian product $F \times CH$. The ideal cognitive agent's initial degree of certainty function Pr is defined on some algebra $\mathcal{A}_W$ over W.

Lewis (1986b)'s Humean supervenience assumption with respect to chance implies the following claim. For every[3] factual world f in F there is *exactly one* possible chance measure ch_f in CH such that (f, ch_f) is in W. Therefore, for any two modal worlds $w = (f_w, ch_w)$ and $w' = (f_{w'}, ch_{w'})$ from W: if $f_w = f_{w'}$, then $ch_w = ch_{w'}$ and, hence, $w = w'$. This means that a modal claim such as $ch_t(A) = x$ amounts to a purely factual claim that is entirely about local matters of particular fact. More generally, it means that the algebra of potentially modal propositions $\mathcal{A}_W$ reduces to the algebra of purely factual propositions $\mathcal{A}_F$.

The latter implies that chances can automatically be iterated indefinitely so that it makes sense to speak of the chance at an earlier time t_0 of the chance at a later time t_1 that some factual proposition A is true, $ch_{t_0}(ch_{t_1}(A) = y) = x$ – something that is also possible, just not automatically so, if the nested character of modality is acknowledged. However, this comes at the expense of losing the distinction between fact and modality: information about chances has become information about local matters of particular fact, and information about local matters of fact has become information about chances. This in turn leads to the problem of undermining futures (Hall 1994, Lewis 1994, Thau 1994).

In a nutshell, the problem is that the principal principle, the assumptions about admissibility, and Humean supervenience with respect to chance together imply that there are no propositions E, possible worlds w, and times t such that: (i) E is a possible future of w at t in the sense that the conjunction of the complete history of w up to t, $H_{w,t}$, and E, $H_{w,t} \cap E$, specifies a possible totality of local matters of particular fact (which is the case if E is a future of w at t that has a non-zero chance in w at t, $ch_{w,t}(E) > 0$); and (ii) this totality $H_{w,t} \cap E$ yields chances $ch_{v,t}$ at t that differ from those in w, $ch_{v,t} \neq ch_{w,t}$.

[3]To simplify the derivation of the problem of undermining futures, I follow Stalnaker (1996: §4) who suggests that Lewis (1986b)'s Humean supervenience assumption should not be viewed as a contingent thesis. If one did, one would have to restrict the scope of the universal quantifier to a proper subset of F (that contains the actual factual world). The problem of undermining futures persists if Humean supervenience is a contingent thesis (see Briggs 2009).

According to Humean supervenience, the possible totality of local matters of particular fact $H_{w,t} \cap E$ has a unique theory of chance T_v, for some possible world $v \in H_{w,t} \cap E$. According to the principal principle and the assumptions about admissibility, T_v and the complete history of (any possible world in) $H_{w,t} \cap E$ up to t together specify the chance distribution $ch_{v,t}$ of v at t. If $ch_{w,t} \neq ch_{v,t}$, then, since (all possible worlds in) $H_{w,t} \cap E$ and w have the same complete history up to t, $T_v \cap T_w = \emptyset$. Since T_v supervenes on $H_{w,t} \cap E$ and Humean supervenience is not a contingent thesis, $H_{w,t} \cap E = H_{w,t} \cap E \cap T_v$. Hence, $H_{w,t} \cap E \cap T_w = H_{w,t} \cap E \cap T_v \cap T_w = \emptyset$. In other words, any merely possible future E of w at t is incompatible with $H_{w,t} \cap T_w$. This implies that $\bigcup_E H_{w,t} \cap E \cap T_w = H_{w,t} \cap E_{w,t} \cap T_w$, where E now ranges over all possible futures of w at t and $E_{w,t}$ is the actual future of w at t. Since $\bigcup_E H_{w,t} \cap E \cap T_w = H_{w,t} \cap T_w$, it follows that $H_{w,t} \cap E_{w,t} \cap T_w = H_{w,t} \cap T_w$.

This means w's theory of chance, T_w, supervenes on not merely the totality of local matters of particular fact of w, i.e., the conjunction $H_{w,t} \cap E_{w,t}$ of the complete history of w up to t, $H_{w,t}$, and the actual future of w at t, $E_{w,t}$. Rather, T_w supervenes on the local matters of particular fact of w up to an arbitrarily early point t in w's history. It also means the chance distribution of w at t, $ch_{w,t}$, supervenes on the local matters of particular fact of w up to t.

In addition, it follows that, for an arbitrarily early point in time t, the future of w at t supervenes on the complete history of w up to t and w's theory of chance. Since w's theory of chance supervenes on the complete history of w up to t, and the future of w at t supervenes on the conjunction of the two, the future of w at t supervenes on the past of w at t, for any possible world w and an arbitrarily early point t in its history. That much supervenience is too much supervenience even for the defender of Humean supervenience.

The problem of undermining futures does not arise if we acknowledge the nested character of modality according to which modal claims qualify factual claims. On the idealist version of this view, Humean supervenience does not do any positive work, but only robs modality of its character. Acknowledging this character exposes Humean supervenience as what it is: an attempt to Humeanize an unHumean theory. Better to not adopt such a theory in the first place.

Finally, while we have restricted the discussion to one modality, modalities can be iterated and mixed. Suppose we want to consider the chance at one time of the chance at another time that some factual proposition is true, or what the chance that some factual proposition is true would have been if things had been a certain way. In this case a modal world takes the form of a triple (f_w, ch_w, ch'_w) and (f_w, ch_w, r_w), respectively, where the possible chance measures ch'_w and the possible atypicality functions r_w are defined on an algebra over $F \times CH$.

7.4 The royal rule

Chances develop over time, and to simplify the discussion, we have granted Lewis (1980)'s assumption that it makes sense to speak of the time in different possible worlds. Normality may also develop over time, but is, more generally, dependent on context (of which the complete history up to a certain time is a special case). Given the truth conditions of counterfactuals from section 7.2, this dependence is illustrated by the following counterfactuals (after Quine 1960: §46; see also Goodman 1954/1983: ch. 1) each of which is said to be true in some context:

K1 If Caesar had been in command in Korea, he would have used only 20th century weapons such as nuclear bombs.

K2 If Caesar had been in command in Korea, he would have used only weapons from the era of the Roman Empire such as catapults.

In the literature (Stalnaker 1999: ch. 1, Gillies 2009) contexts are represented as relations between (centered) worlds or functions from (centered) worlds to sets of such. Since the modal idealist who acknowledges the nested character of modality has a slightly more complicated picture, we represent contexts as functions from modal worlds to sets of factual worlds. This means that what can be presupposed is restricted to the realm of the factual. Formally, a context c is a function from W into $\{F' \times AT : F' \subseteq F\}$ (not $\wp(F) \times \{AT\}$, as incorrectly stated in Huber 2014). Here, $W \subseteq F \times AT$ is the set of all modal worlds $w = (f_w, r_w)$, and AT is the set of all atypicality functions r_w that are defined on some algebra $\mathcal{A}_F$ over the set F of all factual worlds f_w (unless noted otherwise, this algebra is the powerset of F and $W = F \times AT$). c is *alethically respectable* if, and only if, for all w in W: $w \in c(w)$. In contrast to Gillies (2009), Stalnaker (1998) argues against this assumption as a definitional constraint on all contexts. I agree: not every context is alethically respectable. However, those contexts in which an ideal doxastic agent aims at believing the truth and shunning error, and in which the royal rule is supposed to hold, are. Stalnaker (1998) merely requires all contexts c to be *deontically respectable* in the sense that for all w in W: $c(w) \neq \emptyset$. Both Gillies (2009) and Stalnaker (1998) assume further that all contexts c are *autodoxastically respectable* in the sense that for all w and w' in W: $c(w) \subseteq c(w')$ if $w' \in c(w)$.[4] I do not assume so.

[4]Strictly speaking, Stalnaker (1998) represents contexts by a possible worlds semantics with an accessibility relation between worlds that is serial, as well as transitive and Euclidian, but not necessarily reflexive. These requirements translate into deontic (serial), autodoxastic (transitive plus Euclidian), and alethic (reflexive) respectability.

Royal Rule. *An ideal doxastic agent's grade of disbelief function R should be a regular and completely minimitive ranking function on the powerset* $\wp(F \times AT)$ *of* $F \times AT$ *satisfying the difference formula such that for all alethically respectable contexts c, all numbers n, all factual propositions* $A \in \mathcal{A}_F$*, and all propositions* $E_{A,c} \subseteq F \times AT$ *that are admissible for A in c:*

$$R(A \times AT \mid r_c(A) = n \cap E_{A,c}) = n$$

Here, $r_c(A) = n$ *is the proposition that the degree of atypicality in c that A is true exists and equals n.*

The royal rule is a variant of the principal principle. While the latter relates chance and certainty, the former relates normality and belief. Both principles are instances of the idea that, in the absence of information that is not admissible, alethic modality – whether it is deterministic or indeterministic – constrains or guides doxastic modality. The principal principle relates chance and conditional degree of certainty; the royal rule relates degree of atypicality and conditional grade of disbelief. Specifically, the royal rule says that an ideal doxastic agent's grade of conditional disbelief in a factual proposition A given that, in context c, it is atypical to degree n for A to be true, and no further information that is not admissible for A in c, should equal n.

To illustrate, suppose you are certain of nothing but the tautology. Then you should not conditionally believe that it will not rain tomorrow given that it is not atypical for it to rain tomorrow. However, given that it is atypical for it to rain tomorrow, you should conditionally believe that it will not rain tomorrow. In addition, your grade of conditional disbelief that it will rain tomorrow should be higher, the more atypical it is for it to rain tomorrow according to the information you conditionalize on. Of course, you should not conditionally believe that it will not rain tomorrow given that it is atypical for it to rain romorrow *and* that it will rain tomorrow. The latter information overrides the former; the former fails to screen off the latter.

So far I have followed Lewis (1980) in treating admissibility as a primitive that is sufficient, but perhaps not necessary, for conditional doxastic independence. However, in order to justify the royal rule, as we will attempt in section 8.1, we will follow Spohn (2010) and identify the two. As a consequence, this notion is relative to the chance proposition $ch_t(A) = x$ (alternatively: the atypicality proposition $r_c(A) = n$) rather than (merely) time t (alternatively: context c) and target proposition A. In addition, we will turn assumptions about admissibility into requirements on conditional doxastic independence.

Chance develops over time, and Lewis (1980) assumes historical information about times no later than t to be admissible at t for all propositions. Atypicality depends on context, and we can assume accordingly that the (intersection of all) presupposition(s) of world w in context c, $c(w)$, is admissible in c. However, this is just to make explicit the presuppositions of a conversational context and to acknowledge the latter. It is not required for the means-end argument to follow. For this argument, no non-tautological factual proposition needs to be admissible, and the reference to the conversational context can be dropped.

The second assumption Lewis (1980) makes is that information about how chances at any time t depend on complete histories up to t is admissible at all times for all propositions. The background for this assumption is Lewis (1980)'s identification of a possible world w's theory of chance with the conjunction of all history-to-chance conditionals that are true at w. Without this identification, the assumption amounts to the idea that information that is *entirely* about a possible world's theory of chance is admissible at all times for all propositions. More generally, the idea is that *purely* modal information is admissible. (Note that, on Lewis 1980's theory, the proposition $ch_t(A) = x$ is not, in general, purely modal.)

We turn this assumption about the admissibility of purely modal information into the following requirement: for every alethically respectable context c and every factual proposition A, the ideal doxastic agent should consider every purely modal proposition E that is consistent with the atypicality proposition $r_c(A) = n$ to be conditionally independent of $A \times AT$ given this atypicality proposition. Here, a proposition $E \subseteq F \times AT$ is purely modal if, and only if, if $(f, r) \in E$ for some factual world $f \in F$ and some atypicality function $r \in AT$, then $(f', r) \in E$ for all factual worlds $f' \in F$. This is all we require of admissibility – and all we assume of it, since we do not assume anything of it. Furthermore, this requirement will be covered by the attempted justification of the royal rule in section 8.1.

For the modal idealist who acknowledges the nested character of modality, there is no need to come up with a factual proposition over F that is the theory of modality of modal world w. That would be misunderstanding this character. Rather, the theory of modality of modal world w simply is what it is: in the case of chance it is ch_w; in the case of atypicality it is r_w. The corresponding propositions over $F \times CH$ and $F \times AT$ are $T_w = F \times \{ch_w\}$ and $D_w = F \times \{r_w\}$, respectively. Since T_w is not a proposition over F, it is not in the domain of any chance distribution $ch_{v,t}$, as Hall (1994) would have it (see also Vranas 2004) and Hoefer (1997) criticizes. Information about chances and information about local matters of particular fact are kept separate, and the problem of undermining futures does not arise. In the same way, D_w is not in the domain of any $r_{v,c}$.

The presuppositions of possible world w in context c, $c(w)$, and w's theory of atypicality, D_w, combine to the conjunction $c(w) \cap D_w$. In alethically respectable contexts, this conjunction is never empty because $(f_w, r_w) \in c(w) \cap D_w$. Since R is regular – which, in contrast to the probabilistic case, is always possible – this implies that $R(c(w) \cap D_w) < \infty$. Hence, $R(A \times AT \mid c(w) \cap D_w)$ is well-defined for every factual proposition $A \in \mathcal{A}_F$.

Furthermore, $c(w)$ and D_w together imply, for every factual proposition A, that $r_c(A)$ is what it is in modal world w in context c: $r_c(A) = r_{w,c}(A)$. Consequently,

$$c(w) \cap D_w = (r_c(A) = r_{w,c}(A)) \cap c(w) \cap D_w.$$

Since $c(w) \cap D_w$ is consistent with $r_c(A) = r_{w,c}(A)$, the agent should consider it to be conditionally independent of $A \times AT$ given $r_c(A) = r_{w,c}(A)$. The royal rule then implies that, for every factual proposition A,

$$r_{w,c}(A) = R(A \times AT \mid c(w) \cap D_w).$$

In words: the atypicality distribution of any modal world w in any alethically respectable context c, $r_{w,c}$, equals an ideal doxastic agent's conditional grade of disbelief function whose condition is the conjunction of all presuppositions of the conversational context c and w's theory of atypicality. As we have seen in chapter 5, the conditional grade of disbelief function an agent should have, given certain ends of hers, is a regular and completely minimitive ranking function on the powerset of W satisfying the difference formula. As we will see in section 8.1, given further ends of hers, she should also obey the royal rule. Therefore, the metaphysical thesis that degrees of atypicality are completely minimitive ranks is a necessary condition for the possibility of attaining these ends.

Furthermore, if we assume counterfactuals and default conditionals to express propositions that are true or false at possible worlds, as well as to have the truth conditions stated in section 7.2, we arrive – with the help of results in Raidl (2019) – at the following conclusion: the metaphysical thesis that non-iterated default conditionals satisfy the logical postulates of **VT**, and non-iterated counterfactuals those of **VW**, is a necessary condition for the possibility of attaining these ends.

The three assumptions from section 7.2 are not needed for the derivation of the metaphysical thesis that atypicality has the structure of a ranking function. For this thesis to follow from the royal rule we merely need to assume, as I do, that normality is a matter of degree: a quantitative or numerical concept. That degrees of atypicality are extended natural numbers already is a consequence of the royal rule, not anymore one of its assumptions (see sections 4.1 and 5.1).

For the sake of theoretical generality, I have continued to include the reference to the conversational context c. However, for the above means-end arguments to go through, it is enough to consider the "philosophical" context c^* where, relative to $F \times AT$, nothing is presupposed: $c^*(w) = F \times AT$ for every modal world $w \in F \times AT$. Finally, while we will attempt to justify the royal rule in its entirety, it is worth noting that one single case in which it holds suffices for our means-end arguments. The case in which the ideal doxastic agent suspends judgment about every contingent factual proposition, and in which she obeys the royal rule and lets atypicality guide her conditional grades of disbelief, suffices to fix the logical properties of normality and, given our three assumptions, the logical postulates satisfied by default conditionals and counterfactuals.

The latter are repeated here from section 5.5. Let $>$ be a generic conditional in a propositional logic (over a formal language $\mathcal{L}$) that may be embedded, but not iterated. It follows from the results in Raidl (2019) that the logic of non-iterated default conditionals is the system **VT** that is characterized by 1.-10.

1. $\alpha, \alpha \rightarrow \gamma \vdash \gamma$
2. From $\vdash \alpha \leftrightarrow \beta$ infer $\vdash (\alpha > \gamma) \leftrightarrow (\beta > \gamma)$
3. From $\vdash \beta \rightarrow \gamma$ infer $\vdash (\alpha > \beta) \rightarrow (\alpha > \gamma)$
4. $\vdash \alpha$ if α is a truth-functional tautology
5. $\vdash \alpha > \alpha$
6. $\vdash (\alpha > \beta) \wedge (\alpha > \gamma) \rightarrow (\alpha > (\beta \wedge \gamma))$
7. $\vdash (\alpha > \gamma) \wedge (\beta > \gamma) \rightarrow ((\alpha \vee \beta) > \gamma)$
8. $\vdash (\alpha > \beta) \wedge (\beta > \alpha) \rightarrow ((\alpha > \gamma) \leftrightarrow (\beta > \gamma))$
9. $\vdash (\alpha > \gamma) \wedge \neg (\alpha > \neg\beta) \rightarrow ((\alpha \wedge \beta) > \gamma)$
10. $\vdash (\alpha > \bot) \rightarrow \neg\alpha$

Furthermore, it follows from the results in Raidl (2019) that the logic of non-iterated counterfactuals is the system **VW** that is characterized by 1.-10^+., where 10^+. is the following strengthening of 10. known as modus ponens for $>$.

10^+. $\vdash (\alpha > \gamma) \rightarrow (\alpha \rightarrow \gamma)$

These claims depend on the following supplementary truth condition for (default) conditionals whose condition is infinitely atypical: if the degree of atypicality of the proposition expressed by α at the world of evaluation is infinite, then $\alpha > \gamma$ is true at this world if, and only if, γ is true in all worlds in which α is true.

This is the supplementary truth condition adopted for all conditionals in this book and the first volume. It follows from the truth conditions for counterfactuals, but not those for default conditionals, stated in section 7.2. However, it conflicts with the supplementary truth condition traditionally adopted. Modal logicians typically state truth conditions in terms of accessibility: a world v is accessible from a world w if, and only if, every proposition that contains v is assigned a finite degree by w's atypicality function. Here and elsewhere we identify the degree of atypicality of a proposition $E \subseteq F \times AT$ with the degree of atypicality of its factual component $fct(E) = \{f \in F : \exists r \in AT((f,r) \in E)\}$. Traditionally, conditionals whose conditions are inaccessible from a world – i.e., have an infinite degree of atypicality at this world – are said to be true at this world. On this alternative, traditional supplementary truth condition, 10. fails. However, it is recovered if the atypicality function of every world is regular. The latter is a consequence of the royal rule if, and only if, the context is the philosophical context c^*.

For counterfactuals, the traditional supplementary truth condition conflicts with the one adopted in this book and the first volume. This is so because the latter is a special case of our truth conditions for counterfactuals, which are well-defined even if the antecedent is inaccessible or outright inconsistent, a point I have put misleadingly in section 5.5. This means that we cannot define necessity and possibility in terms of the counterfactual in the usual way. However, as hinted at in section 7.2, we can still give a semantics for necessity and possibility claims, as well as non-conditional normality claims. Necessity is truth in all accessible worlds, possibility is truth in some accessible world, and normality is truth at all worlds with zero degree of atypicality. In contrast to necessity and possibility, normality can also be defined directly in terms of default conditionals in the exact same way as non-conditional belief can be defined directly in terms of conditional belief: normally, α if, and only if, normally, if $\top$, then α.

The royal rule implies that the logical properties of necessity and possibility are precisely the ones that are determined by the properties of context c (for an overview see Garson 2018). This is the reason for the terminology of contexts at the beginning of this section. The logical properties of normality are precisely those of non-conditional belief. However, just like counterfactuals and default conditionals, necessity, possibility, and normality claims cannot automatically be iterated or mixed with other modal claims.

The way the truth conditions of default conditionals and counterfactuals in section 7.2, as well as those of necessity, possibility, and normality claims in the previous paragraphs, are stated ignores that a modal world consists of a factual component and a modal component. This is intended: I wanted to adhere to the conventional way of stating truth conditions for modal claims without appealing to the nested character of modality before introducing the latter. However, this requires that one identify the degree of atypicality of a proposition $E \subseteq F \times AT$ with the degree of atypicality of its factual component $fct(E)$. If one does not want to make this identification, these truth conditions can be stated as follows (see Huber 2017: 1572f). For Boolean sentences α and γ whose $[\![\,]\!]$-interpretation are propositions $[\![\alpha]\!]$ and $[\![\gamma]\!]$ of the form $A \times AT$ and $C \times AT$, for some $A, C \in \mathcal{A}_F$, respectively, the default conditional $\alpha \Rightarrow \gamma$ is true at $w \in F \times AT$ if, and only if, all most r_w-typical factual worlds in the factual component A of $[\![\alpha]\!]$ are elements of the factual component C of $[\![\gamma]\!]$. If, in addition, all factual worlds in A that are at least as r_w-typical as the factual component f_w of w are elements of C, then, and then only, the counterfactual $\alpha \,\boxright\, \gamma$ is true at w. Here and elsewhere, the degree of r_w-atypicality of a factual world f is the smallest number n such that $r_w(A) \leq n$ for every factual proposition $A \in \mathcal{A}_F$ with $f \in A$.

While Boolean sentences α and γ express, in essence, factual propositions over F, the counterfactual $\alpha \,\boxright\, \gamma$ expresses a modal proposition over $W \subseteq F \times AT$. However, it does not automatically make sense to iterate the counterfactual. The reason is that this requires there to be second-order atypicality functions r_2 defined on some algebra $\mathcal{A}_W$ over the set of all pairs $(f, r) \in W$ of factual worlds f and first-order atypicality functions r. For two iterations of the counterfactual there need to be third-order atypicality functions r_3 defined on some algebra over a set of triples $(w, r, r_2) \in F \times AT \times AT_2$ of factual worlds w, first-order atypicality functions r, and second-order atypicality functions r_2. Etc. The same is true for the default conditional $\alpha \Rightarrow \gamma$ (and necessity, possibility, and normality claims).

This implies that the non-iterability of some conditional operator $>$ does not imply that $\alpha > \gamma$ does not express a proposition. This blocks the following fairly prominent argument in the literature on – mainly indicative, but in the case of Edgington (2008) also counterfactual – conditionals (Bennett 2003: ch.s 7 and 16, Edgington 1995; 2008, Gibbard 1981). With reference to natural language it is claimed that it is hard if not impossible to iterate (indicative) conditionals. This claim is then taken to be the premise of an argument for the thesis that (indicative) conditionals do not have truth values. The present account shows this argument to be invalid. The purported non-iterability of a conditional can have as reason also the nested character of modality.

According to our means-end argument, the logic of counterfactuals turns out to be the familiar logic of the similarity approach to counterfactuals, but without Stalnaker (1968)'s controversial law of conditional excluded middle (section 7.1) and without Lewis (1973a)'s controversial centering axiom 12. (for criticism see Nozick 1981, Iatridou 2000, Gunderson 2004, Menzies 2004, Leitgeb 2012a; b).

12. $(\alpha \wedge \gamma) \rightarrow (\alpha \mathrel{\Box\!\!\rightarrow} \gamma)$

Bracketing counterfactuals with inaccessible antecedents, as well as helping ourselves towards the limit assumption (which is discussed in the paragraphs to follow and does not affect the logic of counterfactuals, as Lewis 1973a: 121 notes), Lewis (1973a)'s truth conditions for counterfactuals are ours for default conditionals – except that, for Lewis (1973a), the central semantic ingredient is comparative overall similarity among worlds rather than normality at a world, as well as that counterfactuals can be iterated indefinitely. On these truth conditions, syntactic modus ponens for $\Box\!\!\rightarrow$ translates into the semantic principle of weak centering for similarity: each world is at least as similar to itself as every world. This principle is the reason Lewis (1973a)'s truth conditions for counterfactuals have an equivalent formulation in ours for counterfactuals (again, with similarity substituted for normality). Our truth conditions for default conditionals and for counterfactuals coincide if their common semantic ingredient is weakly centered. Consequently, truth conditions different from ours and a semantic ingredient with different properties can determine the same logical postulates for counterfactuals.

On Lewis (1973a)'s truth conditions, the syntactic centering axiom translates into the semantic principle of strong centering for similarity: each world is more similar to itself than any other world. To the extent that one considers the latter semantic principle to be more plausible than the syntactic centering axiom, this suggests that the semantic ingredient to be plugged into Lewis (1973a)'s truth conditions to get the logic of counterfactuals differs from the (ordinary) notion of similarity (or that these truth conditions are not quite right).

This is also the lesson to be be drawn from Fine (1975)'s "future similarity objection," which considers the following counterfactual.

F If Nixon had pressed the button, there would have been a nuclear holocaust.

While F "is true or can be imagined to be so" (Fine 1975: 358), it comes out false if similarity is taken (too) literally – just as any counterfactual comes out false that says that the future would be very dissimilar to what it actually is if the past or present had been different (see also Tichý 1976).

Lewis (1979: 472)'s response, his "system of weights or priorities" governing comparative overall similarity among worlds (to be discussed in section 8.3 and chapter 9), renders F true. However, I think it is fair to say that, in doing so, it also makes clear that the notion of comparative overall similarity among worlds is at best similar to the ordinary notion of similarity.

It may matter little why Lewis terms the central semantic ingredient of his truth conditions similarity. Furthermore, I am not aware of any work discussing this question. So, I can only speculate. That said, one reason might have been conceptual economy – similar to our argument from Occam's razor in section 7.2. Considering the role perceptual similarity plays in Quine (1960) (and, later, Quine 1974), Teilähnlichkeit and Erinnerungsähnlichkeit play in Carnap (1928), and resemblance plays in Hume (1739; 1748), Lewis might have thought that the concept of similarity is one we need on independent grounds (Lewis 1973a: 1). Better to make do with something needed anyways than to postulate something not needed otherwise (a feature of Leitgeb 2012a; b's semantics).

While the royal rule implies that the atypicality function of every world is regular only if the context is the philosophical context c^*, it does imply the limit assumption (Lewis 1973a: 19ff) for normality at a world: for each modal world $w \in W$ and each factual proposition $A \in \mathcal{A}_F$ there is a factual world $f \in A$ such that $r_w(A^-) \leq r_w(A)$ for all factual propositions $A^- \in \mathcal{A}_F$ with $f \in A^- \subseteq A$. For similarity, the limit assumption says that, for every world and every antecedent from $\mathcal{L}^+$ accessible from it, there is an antecedent world that is most similar to it, where $\mathcal{L}^+$ results from $\mathcal{L}$ by allowing for iterations of $\Box\!\!\rightarrow$. Lewis (1973a: 20) argues against it by considering the counterfactual supposition that a line of less than an inch were more than an inch and suggesting that there is no positive real number x such that, in this case, the line would be at least $1 + x$ inches.

As before, while it is perhaps plausible that no line of more than an inch is at least as similar to a line of less than an inch than every line of more than an inch, I find the counterfactuals themselves considerably less plausible than the similarity claims they amount to on Lewis (1973a)'s semantics. The idea that a line could be more than an inch without being at least $1+x$ inches, for some positive real number x, is a mathematical inconsistency. In fact, as Herzberger (1979) shows, on Lewis (1973a)'s truth conditions, the limit assumption for similarity is equivalent to the following counterfactual consistency condition: for all sentences α from $\mathcal{L}^+$ and worlds w, if α is true at some world that is accessible from w, then so are jointly all sentences in the counterfactual theory of α, $T(\alpha) = \{\gamma \in \mathcal{L}^+ : w \models \alpha \Box\!\!\rightarrow \gamma\}$. According to counterfactual consistency, all that would be the case if something possible were to obtain is itself jointly possible. It holds on our truth conditions.

As an aside, there is a curious tension – noted by Amna Zulfiqar (personal correspondence in 2019) – between Lewis (1973a: 20)'s rejection of the limit assumption for comparative overall similarity between possible worlds and Lewis (1968; 1973a)'s claims about the similarity relation between things in counterpart theory. The latter appears to be assumed to satisfy a limit assumption. According to Lewis (1973a: 39), "something has for *counterparts* at a given world those things existing there that resemble it closely enough [...], and that resemble it no less closely than do other things existing there." According to Lewis (1968: 114), something's counterparts resemble it even "more closely than do the other things in their worlds." See Swanson (2014) for further "limit-assuming theories."

Another way to see that in this example, too, counterfactuals and the similarity claims they amount to on Lewis (1973a)'s semantics come apart is to follow Hájek (ms) and turn the example on its head. Presumably, a line of exactly an inch is at least as similar to a line of less than an inch as every line of at least an inch. However, according to Hájek (ms), the following counterfactual is, like most counterfactuals, false: if the line were at least an inch, it would be exactly an inch.

Whether Hájek's thesis (see also Hájek 2021) that most counterfactuals are false is true depends – on the present approach – on the conversational context, as well as which counterfactuals are at issue: causal (interventionist), backtracking, or spurious (acausal). While nothing in this book or the first volume presupposes either the truth or else the falsity of Hájek's thesis, it parallels the situation of knowledge (see section 2.1) if we are speaking strictly. The latter may mean that the conversational context is the philosophical context c^*.

For the counterfactual $\alpha \mathrel{\square\!\!\rightarrow} \gamma$ to be true at the actual world, γ needs to be true at all worlds in which α is true and which are minimally atypical from the point of view of the actual world. In addition, γ needs to be true at all worlds in which α is true and which are at least as typical as the actual world, from its point of view. The latter clause is empty, if the actual world is minimally atypical from its point of view. However, it becomes more demanding the more atypical the actual world is from its point of view. In the limiting case where the actual world is, from its point of view, atypical beyond any finite degree, γ needs to be true in all worlds in which α is true. In this case the counterfactual $\alpha \mathrel{\square\!\!\rightarrow} \gamma$ is stricter even than the strict conditional $\square(\alpha \rightarrow \gamma)$ because, unlike the former, the latter is compatible with inaccessible worlds where α is true and γ is false.

A sufficient condition for the truth of Hájek's thesis is that the actual world is sufficiently atypical from its point of view, where the meaning of 'most' in Hájek's thesis informs the meaning of 'sufficiently.'

A sufficient condition for the falsity of Hájek's thesis is that atypicality at the actual world is sufficiently fine-grained, and that the actual world is not too atypical from its point of view. In this case the antecedent worlds in which a consequent allegedly fails may be argued to be more atypical after all than the actual factual world, albeit so slightly that one easily ignores it. Whichever view, if any, of Hájek's thesis the reader holds, the present approach is content with providing the framework in which to argue for this view.

To conclude this series of examples about which readers are asked to form their own intuitions, let us check if concerns parallel to those of Fine (1975) for the similarity approach arise for the normality approach to counterfactuals. Consider:

S If Ida hosted a party, it would be unlike anything else.

If similarity is taken (too) literally, then, on the similarity approach, S says that even the worlds most similar (to the actual world) in which Ida hosts a party are unlike the actual world. So, if S is true, as it can presumably be imagined to be, weak and strong centering for similarity fail. Now consider:

A If Ida hosted a party, it would be an atypical event.

On the normality approach (and departing from the conventional way of stating truth conditions of modal claims), A says that even those worlds in which Ida hosts a party and whose factual components are most typical (from the point of view of the actual world) have factual components that are atypical from the point of view of the actual world. In particular, the actual factual world itself is atypical, from the point of view of the actual world. I believe this to be the case, at least in the appropriate context. The actual factual world need not be, and is not, the most typical factual world from the point of view of the actual world. Abnormal things happen, just as low chance events occur. Similarly, a line of exactly an inch will not be at least as typical, relative to a line of less than an inch, as every line of at least an inch. And in the appropriate context, the worlds where Nixon presses the button and whose factual components are most typical, from the point of view of the actual world, can presumably all be imagined to be worlds where a nuclear holocaust occurs.

There may be other examples that some readers intuit to be counterexamples to the normality approach to counterfactuals. If so, they reject our assumptions or they do not have the ends the royal rule is a means to attaining. Neither is a problem for means-end philosophy.

Chapter 8

Applications in Metaphysics

In this chapter we will first see that following the royal rule is a means to attaining the end of typically shunning error. This completes the means-end argument for the logical postulates satisfied by default conditionals and counterfactuals: given our assumptions, these postulates are necessary conditions for the possibility of attaining an end from epistemology. Then I will apply the rank-theoretic account of counterfactuals to two problems in metaphysics and the philosophy of science: the counterfactual analysis of actual causation, and the Arrovian impossibility of aggregating comparative aspects of similarity to comparative overall similarity among possible worlds. This chapter heavily relies on Kroedel & Huber (2013) and Huber (2011; 2012; 2016; 2017).

8.1 Why follow the royal rule?

As always, the answer is: because doing so is a means to attaining an end the agent may or may not have. Specifically, the royal rule will be seen to be a necessary and sufficient means to attaining an end that can be thought of as a variant of what James (1896: sct. VII) calls "our first and great commandments as would-be knowers":

> "Believe truth! Shun error!"

The variant results from the following descriptive versions of these imperatives (see section 2.2) that are qualified by normality.

SEBD_A^A For a *fixed* factual proposition $A \subseteq F$: normally, if A is true, then the ideal doxastic agent does not disbelieve A in the sense that $R(A \times AT) = 0$.

BTBD^A_A For a *fixed* factual proposition $A \subseteq F$: normally, if A is true, then the ideal doxastic agent believes A in the sense that $R((F \setminus A) \times AT) > 0$.

SEBD stands for 'shunning error by default' and BTBD stands for 'believing (the) truth by default.' The meaning of these two default conditionals depends on an atypicality function. As we will see, this atypicality function is not one and the same for each antecedent A. Instead, there will be different atypicality functions for different antecedents. While we will not do so, in principle, one can consider whether, normally, if A is true, then the agent does not disbelieve C, SEBD^C_A. This is the reason for flagging these conditions with both a subscript and a superscript.

We will first see that certain agents attain SEBD^A_A, but fail to attain BTBD^A_A, even if the royal rule is obeyed and we make a certain uniqueness assumption (that I misleadingly termed "Stalnaker's assumption" in Huber 2017 and falsely claimed to validate the syntactic law of conditional excluded middle from section 7.1). Next we will see that any agent who obeys the royal rule attains a variant of the former end (and the latter, if we make the uniqueness assumption) that is formulated in terms of conditional belief. Then I will reformulate this variant in order to motivate a final version of this end that will be the official definition of typically shunning error. We will see that, for a fixed factual proposition A, this end is attained for A by all and only these agents who obey the royal rule for A. This will conclude my argument for the thesis that obeying the royal rule is a necessary and sufficient means to attaining the end of typically shunning error.

Consider a consistent factual proposition $A \subseteq F$ and its complement $F \setminus A$. There are exactly two possibilities: A is true and $F \setminus A$ is false, or it is the other way round. For each of these two possibilities there are exactly three possibilities: both A and $F \setminus A$ are minimally atypical; A is, and $F \setminus A$ is not, minimally atypical; A is not, and $F \setminus A$ is, minimally atypical. This leaves us with six possibilities so far. For each of these six possibilities there are exactly three possibilities: neither A nor $F \setminus A$ is believed; F is not, and $F \setminus A$, is believed; A is, and $F \setminus A$ is not, believed. We thus end up with a total of 18 sets of cases to consider.

Let us first assume that the agent is *modally agnostic* and suspends judgment about whether any contingent factual proposition A is atypical: $R(r(A) = n) = 0$ and $R(r(F \setminus A) = n) = 0$ for all extended natural numbers n from $\mathbb{N} \cup \{\infty\}$. More generally, an agent with grade of disbelief function R is modally agnostic if, and only if, $R(F \times \{r\}) = 0$ for each atypicality function $r \in AT$.

What the royal rule asks of modally agnostic agents can perhaps be motivated by analogy to Moore (1942)'s paradox. The latter is exemplified by sentences of the form 'A and I do not believe that A' or, better suited for our purposes:

A and I disbelieve that A.

Moorean sentences are consistent, but allegedly odd to be believed or asserted. Disbeliefs come in grades, though. So, we arrive at graded versions of Moorean sentences:

A and I disbelieve A to grade n.

The oddity of such graded Moorean sentences increases with n. Minimal oddity for $n = 0$, which just means: A and I do not disbelieve A. Some oddity for $n > 0$, which means: A and I believe that A is false with firmness $n > 0$. Maximal oddity for $n = \infty$, which means: A and I am certain that A is false in the sense that I would never give up my belief that A is false (see the conditional theory of conditional belief from section 5.2).

The last step is to replace grades of disbelief with degrees of atypicality:

A and the degree of atypicality of A equals n.

These sentences are consistent (unless normality is weakly centered, which it is not – section 7.4). They may, or may not, be odd to be asserted. However, they should be disbelieved according to the royal rule; and they should be disbelieved the firmer, the greater the number n. What the royal rule asks of modally agnostic agents is that the conjunction 'A and $r(A) = n$' be disbelieved to grade n.

Agents that are not modally agnostic are additionally asked to add to n their grade of disbelief that $r(A) = n$: 'A and $r(A) = n$' be disbelieved to grade $n + k$, where k is the agent's grade of disbelief that $r(A) = n$. In classical Moorean sentences this additional part is a meta-disbelief about one's first-order disbelief in A. If made explicit, it gives rise to infinitely long Moorean sentences: A and I do not believe that A and I do not believe that I do not believe that A etc. The auto-epistemological reflection principle (section 5.5) requires the agent to be certain of what her own grades of disbelief are. Therefore, it requires all these Moorean sentences to be disbelieved, including the original one.

Let us return to our 18 sets of cases. A case ω consists of a factual component f_ω specifying the truth values of all factual propositions; an alethically modal component r_ω specifying the degrees of atypicality of all factual propositions (as well as the truth values of all default conditionals and counterfactuals with factual antecedents and consequents, if our assumptions hold); and a doxastically modal component R_ω specifying the agent's grades of disbelief for all propositions over $F \times AT$. It can be represented as $\omega = (f_\omega, r_\omega, R_\omega)$ or, stressing the relevant details:

$$\omega = \left(f_\omega(A), r_\omega(A), r_\omega(F \setminus A), R_\omega(A \times AT), R_\omega((F \setminus A) \times AT)\right)$$

In order to make sense of SEBD^A_A and BTBD^A_A we now need to assign degrees of atypicality to these cases ω. Otherwise, the English default conditionals in them do not have a truth value. This is a tricky task because there is the threat that I am smuggling into these degrees of atypicality whatever it is that I want to derive. So, let me try to be as clear as I can. We are given a set of possible cases $\Omega = F \times AT \times S$, where S is the set of all disbelief functions (defined on the power-set of $F \times AT$). Our task is to define one or more atypicality function(s) ϱ on the power-set of Ω.

The antecedents, but not the consequents, of the default conditionals we are considering are restricted to factual propositions. The latter are assigned degrees of atypicality in each case ω, but the cases themselves are not. Yet, in order to evaluate these default conditionals, what we need are degrees of atypicality for the cases themselves, not just their factual components. The reason is that only the cases determine what the agent believes and, hence, whether the consequents of our default conditionals are true. However, since the antecedents *are* restricted to factual propositions, I want to use the alethically modal information about these factual propositions that is contained in the atypicality function r_ω to determine the degree of ϱ-atypicality of a case ω in a bootstrapping manner.

One option is to identify $\varrho_\omega(\{\omega'\})$, for arbitrary cases ω and ω', with $r_\omega(\{f_{\omega'}\})$, as we have done in section 7.4 when stating the truth conditions of modal claims in the conventional way that fails to acknowledge the nested character of modality. However, this works only if not only the antecedents of the default conditionals, but also their consequents are restricted to factual propositions.

Another option is to identify $\varrho(\{\omega\})$ with $r_\omega(\{f_\omega\})$. This avoids the problem mentioned above and tells us how atypical ω is *qua* f_ω-case, namely atypical to degree $r_\omega(\{f_\omega\})$. However, it does not tell us how atypical ω is *qua* A-case, for an arbitrary factual proposition $A \subseteq F$. This information is not provided by $r_\omega(\{f_\omega\})$. It is provided only by a case ω's atypicality function r_ω *in its entirety*, that is, by the degrees of atypicality $r_\omega(A)$ for *all* factual propositions $A \subseteq F$, not just the particular factual proposition $\{f_\omega\}$.

The option I will take is to make the latter idea more precise, and to stay as neutral as possible and only use as much information from r_ω as is needed in order to evaluate our default conditionals. A consequence will be that the degrees of ϱ-atypicality of a case ω depend on the antecedent A of the default conditional that is evaluated. That is, we will end up with a model of the form: $\left(\Omega, (\varrho_A)_{A \subseteq F}\right)$, where, for each factual proposition $A \subseteq F$, ϱ_A is a ranking function on the power-set of Ω that specifies how atypical a case ω is *qua* case in which A is true.

Suppose, then, two cases ω_1 and ω_2 agree that the factual proposition A is true. Under this assumption, the question we need to answer is this: is ω_1 more typical *qua* A-case than ω_2 in the sense of what I will call the "actual atypicality function" ϱ_A? I propose that this be the case if, and only if, the alethically modal components of ω_1 and ω_2 – r_{ω_1} and r_{ω_2}, respectively – say so of A:

$$f_{\omega_1}(A) = f_{\omega_2}(A) = \mathit{true} \Rightarrow \left[\varrho_A(\{\omega_1\}) < \varrho_A(\{\omega_2\}) \Leftrightarrow r_{\omega_1}(A) < r_{\omega_2}(A)\right]$$

Given this assumption, it turns out that modally agnostic agents who obey the royal rule attain SEBD_A^A, but fail to attain BTBD_A^A spectacularly even if the uniqueness assumption holds. The latter says that for each factual proposition $A \subseteq F$ and case ω exactly one of A and $F \setminus A$ is minimally r_ω-atypical (so that there is exactly one factual world f that is minimally r_ω-typical, but not necessarily the actual factual world f_ω – the latter would mean that normality is strongly centered, which it is not; see section 7.4). The reason is that the royal rule implies that agents who are modally agnostic are also factually agnostic. The agent normally does not believe anything false simply because she never believes anything non-tautological in the first place. So, let us drop the assumption of modal agnosticism and consider the following variants of SEBD_A^A and BTBD_A^A that are formulated in terms of conditional belief (where C stands for 'conditionally').

CSEBD_A^A For a *fixed* factual proposition $A \subseteq F$: normally, if A is true, then the ideal doxastic agent does not conditionally disbelieve A given the truth about the alethically modal status of A in any form, i.e., $R(A \times AT \mid r(A)^*) = 0$.

CBTBD_A^A For a *fixed* factual proposition $A \subseteq F$: normally, if A is true, then the ideal doxastic agent conditionally believes A given the truth about the alethically modal status of $F \setminus A$ in any form, i.e., $R((F \setminus A) \times AT \mid r(F \setminus A)^*) > 0$.

$r_\omega(A)^*$ is *any* purely modal proposition that is true in ω and correctly specifies the degree of atypicality of A as $r_\omega(A)$. Recall: a proposition $E \subseteq F \times AT$ is purely modal if, and only if, if $(f, r) \in E$ for some $f \in F$ and some $r \in AT$, then $(f', r) \in E$ for all $f' \in F$. (The dependence of $r_\omega(A)^*$ on ω will become clear in the reformulation of CSEBD_A^A stated in the following paragraphs.) This means that, for a given case ω, we quantify over *all* purely modal propositions $r_\omega(A)^*$ such that:

$$F \times \{r_\omega\} \subseteq r_\omega(A)^* \subseteq \{v = (f_v, r_v) \in F \times AT : r_v(A) = r_\omega(A)\} = (r(A) = r_\omega(A))$$

Given our assumption, it turns out that agents who obey the royal rule attain CSEBD_A^A, as well as CBTBD_A^A if the uniqueness assumption holds for A.

To motivate an end that is stronger than CSEBD^A_A and that will be my official definition of typically shunning error, consider the following reformulation that makes the dependence of $r_\omega(A)^*$ – and of R_ω – on ω clear:

CSEBD^A_A For a *fixed* factual proposition $A \subseteq F$: a case ω in which A is true and the ideal doxastic agent conditionally disbelieves A given the truth about the alethically modal status of A in ω in some form, $R_\omega(A \times AT \mid r_\omega(A)^*) > 0$, is an atypical case in which A is true, i.e., $\varrho_A(\{\omega\}) > \varrho_A(A \times AT \times S)$.

This reformulation suggests the following strengthening of CSEBD^A_A, which is my official definition of typically shunning error.

TSE^A_A For a *fixed* factual proposition $A \subseteq F$: of two possible cases ω_1 and ω_2 in which A is true, the former is more atypical *qua* A-case than the latter if, and only if, in ω_1 the ideal doxastic agent conditionally disbelieves A given the truth about its alethically modal status in ω_1 (in some form) to a high grade, whereas in ω_2 she (that is, her counterpart R_{ω_2}) conditionally disbelieves A given the truth about its modal status in ω_2 (in some form) to a low, and possibly no, degree, i.e.:

$$R_{\omega_1}(A \times AT \mid r_{\omega_1}(A)^*) > R_{\omega_2}(A \times AT \mid r_{\omega_2}(A)^*) \Leftrightarrow \varrho_A(\{\omega_1\}) > \varrho_A(\{\omega_2\})$$

Theorem 8. *An ideal doxastic agent whose beliefs obey the ranking calculus obeys the royal rule for A if, and only if, she attains TES^A_A. Consequently, she obeys the royal rule if, and only if, she attains TSE^A_A for all factual propositions A.*

PROOF: This follows from the considerations in Huber (2017: 1584ff) plus the stipulation that the degree of atypicality of the empty set $\emptyset$ equals ∞. *Q.E.D.*

Three points are to be noted. First, it does not make sense to strengthen CBTDB^A_A in analogy to TSE^A_A. The reason is that, while (conditional) beliefs and disbeliefs come in grades, (conditional) suspensions of judgment do not.

Second, the royal rule for A includes the requirement that, for every extended natural number n, the agent should consider every purely modal proposition E that is consistent with the atypicality proposition $r(A) = n$ to be conditionally independent of $A \times AT$ given this atypicality proposition.

Third, an agent pursuing any end stronger than TSE^A_A equally has to obey the royal rule. The reason is that, on the instrumentalist position adopted in this book and the first volume, what one ought to do is take the *necessary* means to attaining one's ends. Since obeying the royal rule for A is necessary and sufficient for attaining TSE^A_A, obeying the royal rule is necessary for attaining any stronger end.

Consequently, an agent considering whether to follow the royal rule should consider not (just) how desirable she finds it to typically shun error in the sense of TSE_A^A. First and foremost, she should consider how *un*desirable she finds it to *fail* to typically shun error. The situation here is parallel to that of the consistency argument (chapter 5): it is primarily not the desirability of consistent beliefs, but the undesirability of inconsistent beliefs one needs to evaluate in deciding whether to follow the rules of ranking theory.

To conclude this section, let us illustrate the end of typically shunning error by an example. Ida and Bay wonder whether Tweety can fly. Both correctly believe that normally, if Tweety is a bird, it can fly. For the sake of definiteness let us additionally assume that they have arrived at their belief in this singular default conditional from their belief in the generic default rule that normally birds can fly. However, singular default conditionals do not follow logically from the corresponding generic default rules, just as claims about single-case chances do not follow logically from the corresponding generic frequency claims. (We will study in chapter 10 under which conditions the truth values of default conditionals and counterfactuals can be reliably inferred from statistical information.)

Ida's beliefs additionally obey the royal rule. She holds the conditional belief that Tweety can fly if it is a bird. Bay, on the other hand, does not hold the conditional belief that Tweety can fly if it is a bird. Her beliefs do not obey the royal rule. After some research, the two friends come to believe that Tweety is a bird, which, in fact, it is.

For Ida there is the following benefit of obeying the royal rule. If Tweety can fly, but Ida incorrectly believes otherwise, no matter how weakly, then this is an atypical situation: a situation that normally does not obtain if some situation obtains. In other words, by correctly believing what is normally the case, Ida normally correctly believes what is the case. Not so for Bay. If Tweety can fly, but Bay incorrectly believes otherwise, then this need not be an atypical situation, no matter how firmly Bay holds her belief. Furthermore, if Tweety can fly, but Ida incorrectly believes otherwise, and she does so firmly, then this is a highly atypical situation: a situation that normally does not obtain even if an atypical situation obtains that normally does not obtain if some situation obtains.

What are such situations like that are not merely atypical, but highly so? They are like the following ones. One should not steal. However, given that one steals, one should steal from the rich and not the poor. Ida is a good person. Normally, she does not steal. Furthermore, normally, if Ida steals, she steals from the rich and not the poor. A situation in which Ida steals is atypical. A situation in which Ida steals from the poor and not the rich is highly atypical.

8.2 Actual causation

Causal relations are many and varied. There is the relation of causal relevance between generic properties such as height and weight that is primarily studied in the sciences. There is the relation of causal relevance between singular properties such as Ida's height and her weight. There is the relation of causation between specific generic properties such as a height of 5′5″ and a weight of more than 120lbs. Finally, there is the relation of actual causation between events or similar entities such as facts – or aspects (Paul 2000) or similar entities such as tropes – such as Ida's being 5′5″ tall and her weighing more than 120lbs. These causal relations can, of course, be divided further. For instance, one may distinguish between actual causes that are necessary, sufficient, contributing, direct, etc.

Generic properties can be represented by generic variables, while singular properties can be represented by singular variables. The former are measurable functions whose domain is a population of individuals I from which one can – at least, in principle – draw samples (for measurability see Huber 2018: sct. 10.3). The latter are measurable functions whose domain is a set of mutually exclusive possible worlds W from which one cannot draw samples.

To illustrate the difference between singular and generic variables, let us – temporarily – identify possible worlds with models for a first-order language. These are ordered pairs $\langle D, [\![]\!]\rangle$ consisting of a domain of individuals D and an interpretation function $[\![]\!]$. The domain of a singular variable is (a subset of) the set of all models for the language. By contrast, the domain of a generic variable is a subset of the domain $D_@$ of the actual model $@ = \langle D_@, [\![]\!]_@\rangle$ that accurately and, relative to the language, maximally specifically describes reality.[1]

Specific generic properties can be represented by the taking on of specific values of generic variables: $X = x$ is the specific generic property comprising the set of individuals i that have X to degree x, $\{i \in I : X(i) = x\}$; $X > x$ is the specific generic property comprising the set of individuals i that have X to degree greater than x, $\{i \in I : X(w) > x\}$. Events and aspects (and similar entities such as facts and tropes) can be represented by the taking on of specific values x of singular variables X: $X = x$ is the proposition comprising the set of possible worlds w in which X takes on x, $\{w \in W : X(w) = x\}$; $X > x$ is the proposition comprising the set of possible worlds w in which X takes on values greater than x, $\{w \in W : X(w) > x\}$.

[1] Alternatively, the domain of a generic variable can be taken to be (a subset of) the set of possible individuals in any domain rather than a subset of merely the domain of actual individuals $D_@$. However, samples can be drawn only from the latter.

Theories of actual causation can be classified into three categories: regularity theories, associationist theories, and (for lack of a better term) cognitivist theories. Depending on one's exegesis, accounts in the tradition of Hume (1739; 1748) and Mill (1843) fall into the first or second category, while realist accounts fall into the third. Very crudely, regularity theories aim at making do with regularities of factual matters only. Associationists and cognitivists find these regularities lacking in flexibility. So, they additionally help themselves towards doxastic and alethic modalities, respectively. This does not come without cost, though. The former end up with a partially subjective notion of actual causation, while the latter are (seemingly) less parsimonious in their ontology.

My categorization may not be exhaustive (or exclusive). However, it includes process or conserved quantity theories (Salmon 1984; 1998, Dowe 2000), as well as agency theories (e.g., Menzies and Price 1993). The reason is that, while this is not their distinctive feature, these theories employ probabilities or counterfactuals.

If the theory of actual causation is probabilistic, the three categories reflect the interpretation of probability: regularity theories work with a (limiting) relative frequency interpretation (Reichenbach 1956); associationist theories work with a subjective interpretation (Skyrms 1980); and cognitivist theories work with an interpretation, such as chance, that makes sense of single-case probabilities and is not subjective (Suppes 1970 does not commit to a specific interpretation). If the theory is not probabilistic, the three categories reflect the nature of the central non-probabilistic concept employed: approaches that work within the confines of first-order logic such as Mackie (1965)'s (see also Baumgartner 2013 and Baumgartner & Falk 2021) classify as regularity theory; approaches that work with conditional beliefs such as Spohn (2006)'s classify as associationist; and approaches that work with counterfactuals such as Lewis (1973b)'s classify as cognitivist.

The causal models (Spirtes et al. 1993/2000, Pearl 2000/2009) to which we will turn in the next chapter work with both probabilities and non-probabilistic concepts. They too can be used to formulate theories of actual causation (Halpern 2016). However, the directed graphs (and, if used, structural equations) these models employ do not provide a reductive theory of any causal relation. Rather, they are representations of causal assumptions. So, unlike the accounts mentioned in the previous paragraph, the resulting theories of actual causation do not offer reductive analyses to non-causal concepts. Instead, they analyze actual causation in other causal terms (see Cartwright 1979 for an argument against the possibility of a reductive analysis of actual causation). This is so independently of the fact that the notion of an intervention that is generally used to interpret causal models too is a causal concept (Woodward 2003).

Not all are enthusiastic about actual causation (Glymour et al. 2010). So, it is worth mentioning that Hitchcock & Knobe (2009) argue on empirical grounds that the criteria seemingly governing judgments of actual causation are informed by norms. Furthermore, they argue that these criteria serve a legitimate purpose: they "are designed in such a way that they generally tend to direct us towards strategies of intervention that would be preferable" (612). There are several senses in which a strategy can be preferable. These are determined by different norms: statistical norms, moral norms, and norms of proper functioning. Let us focus on statistical norms. As the authors stress: "[w]hat we will want is a strategy of intervention that is *generalizable* – a strategy that would be effective not just in this one situation but also in other situations of a roughly similar type" (607).

Given the means-end perspective adopted in this book, as well as the first volume, I can only applaud these considerations. That said, in order to provide the justification for the concept of actual causation the authors take their findings to provide, I think their argument needs amendment (analogous remarks apply to Morris et al. ms). The reason is that, whatever else actual causation is, it is a relation between events or aspects or similar entities that can be represented by the taking on of specific values of *singular* variables. Yet "a strategy of intervention that is *generalizable*" requires a *generic* variable. If it really is the selection of an intervention that "would *generally* be a good strategy" (607) that justifies a causal concept, then this concept is the concept of causation between specific generic properties, not the concept of actual causation. Conversely, if it really is the concept of actual causation that serves a legitimate purpose, then it does so not because its criteria of selection point to "strategies of intervention that are generalizable" (608), but because its criteria of selection point to an intervention that is good in the singular sense of descriptive normality from section 7.2 or some other singular, not generic or statistical sense: an intervention that would be preferable in the one situation in which it is or is not carried out.

To illustrate, a couple shares an intimate moment, one partner suffers a cardiac arrest. The former event is an actual cause of the latter. However, "[i]f our goal is to prevent [partners from suffering cardiac arrests], it would *generally* [not] be a good strategy [in any of the three senses of good the authors identify] to make sure" (607) couples do not share intimate moments. It would have been a good intervention on this particular occasion, not generally. The singular sense of descriptive normality may perhaps help deliver this verdict: while it is statistically and morally normal for couples to share intimate moments, as well as normal for couples in properly functioning relationships, the particular intimate moment our couple shared still was special (for reasons other than that it led to a cardiac arrest).

We will return to the distinction between singular and generic variables, and why it matters, in the next chapter.

While the question of which purpose, if any, is served by the concept of actual causation may remain open, this need not undermine the point of this section. The latter is to show that the modal idealist equips the theorist of actual causation with the flexibility that a regularity theory lacks, but without the costs incurred by the associationist and cognitivist. In the case of a probabilistic theory this is achieved by adopting the modal idealist interpretation of single-case probabilities from sections 7.3 and 7.4. In the case of Lewis (1973b)' counterfactual theory we can even improve upon the results, as I will try to show now. (One can, of course, also combine probabilistic and non-probabilistic concepts, as, e.g., in the cognitivist accounts of Lewis 1986c, as well as, on at least one interpretation, Fenton-Glynn 2017.)

Lewis (1973b) assumes actual causation to be a relation between events (Lewis 1986d). First he defines a special case of actual causation (which differs from both the relation of causal relevance between generic properties and the relation of causal relevance between singular properties): event c is causally relevant to event e in possible world w if, and only if, c and e occur in w, and, in w, e would not have occurred if c had not occurred. The latter counterfactual must not be a backtracking (or spurious) one, something Lewis (1979) attempts to define in non-causal terms (and we will discuss in the next section and chapter). On Lewis (1973a)'s semantics for counterfactuals, the former clause implies that, in w, e would have occurred if c had occurred. The reason is that this semantics validates the centering axiom (section 7.4). Actual causation itself then is defined as the transitive closure of causal relevance between events.

If one works with a weaker logical system such as ours, the causal relevance of event c to event e in possible world w has to be defined as follows: c and e occur in w; in w, if c had not occurred, then e would not have occurred; and, in w, if c had occurred, then e would have occurred.

On the normality approach to counterfactuals from chapter 7, the last two clauses imply, but are not implied by, the following inequalities: where C and E are the propositions that c occurs and that e occurs, respectively, and r_w is the typicality function of possible world w, if $r_w\left(\overline{C}\right)$ and $r_w\left(C\right)$ are finite, then C:

$$r_w\left(E \cap \overline{C}\right) > r_w\left(\overline{E} \cap \overline{C}\right) \text{ and } r_w\left(\overline{E} \cap C\right) > r_w\left(E \cap C\right).$$

There are several alleged counterexamples to Lewis (1973b)'s counterfactual theory of actual causation (Collins et al. 2000, Paul & Hall 2013), and Lewis has since refined his account (Lewis 1986c; 2000).

Spohn (1983a; 1983b; 1990; 2006; 2012: ch. 14)'s theory of actual causation can be developed in terms of probabilities, as well as ranking functions. Like Lewis (1973b), Spohn first defines a special case of actual causation termed direct causation (rather than causal relevance between events). Actual causation itself then is again defined as the transitive closure of direct causation.

Unlike Lewis (1973b: 566), who wants to allow that the forward direction of time is defined as the predominant direction of actual causation, Spohn assumes that causes temporally precede their effects. Given this assumption, as well as others that pertain to the formal framework of singular variables of the theory, event c is a direct cause of event e in possible world w if, and only if, c and e occur in w and the following inequality R holds:

$$R\left(\overline{E} \mid C \cap H_w^{c,e}\right) - R\left(E \mid C \cap H_w^{c,e}\right) \;>\; R\left(\overline{E} \mid \overline{C} \cap H_w^{c,e}\right) - R\left(E \mid \overline{C} \cap H_w^{c,e}\right)$$

Here, $H_w^{c,e}$ is the complete history of world w up to right before the effect e, but excluding the cause c (Spohn's formal framework allows a precise formulation of this clause). R is an ideal doxastic agent's grade of disbelief function.

The left-hand side is positive/negative if, and only if, the agent conditionally believes/disbelieves E given C and $H_w^{c,e}$. The right-hand side is positive/negative if, and only if, the agent conditionally believes/disbelieves E given $\overline{C}$ and $H_w^{c,e}$. The inequality says that the agent's conditional belief in E is firmer given C and $H_w^{c,e}$ than given $\overline{C}$ and $H_w^{c,e}$, or her disbelief in E is weaker given C and $H_w^{c,e}$ than given $\overline{C}$ and $H_w^{c,e}$. This can happen in four ways giving rise to supererogatory or additional, insufficient or weak, sufficient, and necessary direct causes, respectively: both sides of the inequality are positive, it is just that the left-hand side is even greater than the right-hand side; both sides are negative, it is just that the right-hand side is even smaller than the left-hand side; the left-hand side is positive, while the right-hand side is not; the right-hand side is negative, while the left-hand side is not.

In the probabilistic case the corresponding inequality is equivalent to

$$\log\left(\frac{\Pr\left(E \mid C \cap H_w^{c,e}\right)}{\Pr\left(\overline{E} \mid C \cap H_w^{c,e}\right)}\right) \;>\; \log\left(\frac{\Pr\left(E \mid \overline{C} \cap H_w^{c,e}\right)}{\Pr\left(\overline{E} \mid \overline{C} \cap H_w^{c,e}\right)}\right),$$

where Pr is a probability measure (that is regular). Here too one could distinguish between four kinds of direct causes by considering whether the two sides of the inequality are positive/negative. However, Spohn (1983a: 210) thinks this should be done only in the rank-theoretic case, perhaps because he considers the simpler inequality P: $\Pr\left(E \mid C \cap H_w^{c,e}\right) > \Pr\left(E \mid \overline{C} \cap H_w^{c,e}\right)$.

In the probabilistic case Spohn eventually (in Spohn 1999; 2010, though not in Spohn 1983a; 1983b; 1990) opts for an interpretation that makes sense of single-case probabilities and is not subjective. Let us call it objective probability. While Spohn (1999; 2010) could understand objective probabilities as projections of subjective degrees of certainty that have no reality themselves, he falls short of taking this ontological stance and restricts himself to epistemological applications. Objective probability remains a primitive. In the rank-theoretic case, Spohn opts for a subjective interpretation in terms of grades of conditional disbelief. This rank-theoretic theory of actual causation can handle the problems besetting Lewis (1973b)'s theory, most notably actual causation by early and late preemption, as well as actual causation by symmetric overdetermination and trumping.

Actual causation by preemption is a case of three occurring events c, d, and e such that (i) c, but not d, is an actual cause of e; (ii) e would not have occurred if neither c nor d had occurred; and, if the centering axiom does not hold, (iii) e would have occurred if c, d, or both had occurred. As far as counterfactuals are concerned, c and d are entirely symmetric. Therefore, one needs to turn to other resources to distinguish the actual cause c from the preempted would-be cause d. In cases of early preemption inserting additional events works. In cases of late preemption fine-graining of existing events works.

Let me try to illustrate these cases with variants of a highly fictitious example. Ida and Bay, the two fastest skiers in the world, both qualify for the Olympics to represent their native Austria, but no other Austrian does (that's the highly fictitious part). For the first variant, only one skier per country may participate in the downhill race. Ida qualified. Bay qualified. Austria wins. Ida's, but not Bay's, qualifying is an actual cause of Austria's winning. If neither Ida nor Bay had qualified, Austria would not have won. If Ida had not qualified, Bay would still have qualified and Austria would still have won. If Bay had not qualified, Ida would still have qualified and Austria would still have won.

To render Ida's, but not Bay's, qualifying an actual cause of Austria's winning we insert an additional event that actually occurred: Ida's being fastest. If Ida had not qualified, she would not have been fastest. If Ida had not been fastest, she would still have started for Austria, she just would not have finished first, so Austria would not have won. However, if Bay had not qualified, Ida would still have qualified, she would still have been fastest, and Austria would still have won. Therefore, while there is a chain of causal relevance between events from Ida's qualifying via her being fastest to Austria's winning, the potential chain of causal relevance between events from Bay's qualifying to Austria's winning never goes to completion: it is cut short by Ida's being fastest.

For the second variant, two skiers per country may participate in the downhill race. Ida finishes in record time. Bay comes in second (but not with her personal best on this course). Austria wins. Ida's, but not Bay's, qualifying is an actual cause of Austria's winning. If neither Ida nor Bay had qualified, Austria would not have won. If Ida had not qualified, Bay would still have qualified and Austria would still have won. If Bay had not qualified, Ida would still have qualified, and Austria would still have won.

To render Ida's, but not Bay's, qualifying an actual cause of Austria's winning we can again try to insert an additional event that actually occurred: Ida's being fastest. If Ida had not qualified, she would not have been fastest. However, this time we do not get the counterfactual that, if Ida had not been fastest, Austria would not have won. For this time Austria would still have won because Bay would have been fastest. So, a different strategy is called for: fine-graining of events. If Ida had not qualified, Austria would still have won. However, it would have done so in a different manner, namely without winning in record time. On the other hand, if Bay had not qualified, Austria would still have won, and it would still have done so in record time. The required asymmetry between the actual cause and the preempted would-be cause has been established.

To the extent that inserting additional events and fine-graining existing events are admissible strategies for some theory of actual causation, they are so for all theories. Lewis (1973b)'s and Spohn's theories of actual causation are on a par with respect to early and late preemption. So, let us turn to actual causation by symmetric overdetermination where the two events c and d should come out as entirely symmetric; as well as actual causation by trumping where c, but not d, is an actual cause of e, but neither inserting nor fine-graining of events works.

For the third variant, two skiers per country may participate in the downhill race. Ida finishes in record time. Bay comes in at the exact same time: *ex aequo*. Austria wins. This is a case of symmetric overdetermination where Ida's and Bay's qualifying are entirely symmetric.

What are the actual causes of Austria's winning? While we will come across more sophisticated approaches in the next chapter, accounts in the tradition of Lewis (1973b) tend to rule that neither Ida's nor Bay's qualifying is (though the disjunctive event, if an event, of Ida's or Bay's qualifying, but not the conjunctive event of their joint qualifying, may be). Spohn's theory renders a different verdict: Ida's and Bay's qualifying (plus their joint and disjunctive qualifyings, if events) both are direct causes. Specifically, each is an additional cause (as is their joint qualifying) in the presence of the other, but would have been a necessary and sufficient cause (as is their disjunctive qualifying) in the absence of the other.

For the fourth variant, everything is as in the second variant except that only the fastest skier of each country counts. This is a case of trumping where Ida's, but not Bay's, qualifying is an actual cause of Austria's winning.

Accounts in the tradition of Lewis (1973b) tend to rule that neither Ida's nor Bay's qualifying is an actual cause of Austria's winning (though the disjunctive event, if an event, of Ida's or Bay's qualifying, but not the conjunctive event of their joint qualifying, may be). Spohn's theory renders a more refined verdict: Ida's (plus their disjunctive, if an event), but not Bay's (or their joint), qualifying is a direct cause. Specifically, Ida's qualifying is an additional cause in the presence of Bay's qualifying, and would have been a necessary and sufficient cause (as is their disjunctive qualifying) in the absence of Bay's qualifying – as would have been Bay's qualifying in the absence of Ida's qualifying.

These verdicts take the agent to hold a conditional belief if the corresponding counterfactual is true. For instance, since Ida would still have qualified if Bay had not, the agent conditionally believes that Ida qualifies given that Bay does not. For the first and second variant, early and late preemption, this is sufficient.

For the third variant, symmetric overdetermination, the agent conditionally believes (as she does in all four variants) that Austria wins given that Ida and Bay qualify; given that Ida, but not Bay, qualifies; and given that Bay, but not Ida, qualifies. Furthermore, she holds the first conditional belief more firmly than both the second and third (which she may, but need not, hold equally firmly). Finally, the agent conditionally believes (as she does in all four variants) that Austria does not win given that neither Ida nor Bay qualifies.

For the fourth variant, trumping, the agent holds the exact same conditional beliefs. However, this time she holds the conditional beliefs that Austria wins given that Ida and Bay qualify and given that Ida, but not Bay, qualifies equally firmly, and more firmly than the conditional belief that Austria wins given that Bay, but not Ida, qualifies. (If she holds the first conditional belief more firmly than the second, and the second more firmly than the third, then we have returned to the third variant, symmetric overdetermination: Bay's qualifying, as well as Ida's and Bay's joint qualifying, are additional causes of Austria's winning.)

Accounts in the tradition of Lewis (1973b) cannot copy these verdicts because counterfactuals, like all propositions, can be only true or false, but not true or false to different degrees. In contrast to this, conditional beliefs can be not only held, but held with different numerical grades of firmness. Also, note that a merely comparative ordering is insufficient to distinguish symmetric overdetermination and trumping: the two ranking functions for the third and fourth variant can be assumed to induce the same implausibility ordering (section 3.3).

Thus, this rank-theoretic theory of actual causation can handle the problems besetting Lewis (1973b)'s theory. However, as Spohn (1983a; 1983b; 1990; 1993; 2012: ch.s 14, 15; 2018) acknowledges, a major problem is that actual causation is rendered as subjective as grade of disbelief. Subjectivity is also an issue of Skyrms (1980)'s probabilistic theory which consists in an inequality similar to P, differing primarily in the constraints imposed on the background condition $H_w^{c,e}$. Skyrms (1980; 1988) addresses it by conditionalizing the subjective degree of certainty function Pr on the true cell of a partition (such as the set of possible complete histories up to right before the effect, but excluding the cause, of which $H_w^{c,e}$ is the true cell), a move already employed by Jeffrey (1965/1983). This strategy is also applicable to ranking functions. However, as Spohn (1993; 2010; 2012: ch. 15) admits, it merely mitigates, but does not solve, the problem.

In fact, conditionalizing on the true cell of a partition is even less promising than Spohn seems to realize. The reason is that one can conditionalize on the true cell of a partition only if one can conditionalize, and one can conditionalize only if there are conditional beliefs and conditional degrees of certainty. What are these, though? Spohn (2012) and Skyrms (1980) tell us merely how they should behave, not what they are (except for a reference to conditional bets in, e.g., Skyrms 1987). The conditional theories of conditional belief and conditional degree of certainty from chapter 5 fill these lacunae. They define these notions in terms of, among others, counterfactuals. Hence, on the theories of chapter 5, any theory of actual causation that works with these notions appeals to counterfactuals. So, any such theory becomes cognitivist on top of being associationist and fails to do without the alethic modalities the associationist considers to be ontologically suspect.

All the better news it is that Spohn (1993; 2012: ch. 15) develops an account of "objectifying" subjective ranking functions. The main thought is as follows: some properties of a subjective ranking function can be brought into one-to-one correspondence with certain non-modal propositions that are "objectively" true or false. For instance, the non-conditional beliefs of a subjective ranking function can be brought into one-to-one correspondence with such non-modal propositions, namely, the contents of those non-conditional beliefs. Hence, the non-conditional beliefs of a subjective ranking function can be objectified: a non-conditional belief is objectively true if, and only if, its content is.

Each property of a subjective ranking function that can be uniquely associated with a non-modal proposition can itself be said to be true or false, depending on whether the associated non-modal proposition is true or false. Thus, the question is which properties of a subjective ranking function can be objectified in this sense and, in particular, whether actual causation can be so objectified.

It turns out that neither direct nor actual causation can be objectified, though both notions, like sufficient, necessary, and necessary and sufficient causation can be "conditionally objectified" (Spohn 1993; 2012: ch. 15). However, the notion at play in the treatment of symmetric overdetermination and trumping is additional causation. Like weak causation, it cannot be objectified with the method from Spohn (1993; 2012: sct. 15.4) – not even conditionally, for conditions which are themselves objectifiable with this method. That said, there is a second method (Spohn 2012: sct. 15.5) with which all of these notions can be conditionally objectified. The condition in the sense of which these notions can be conditionally objectified with one of these methods is that direct causes immediately precede their effects. This in turn presupposes that the singular variables in the formal framework are linearly ordered by time (the subjective theory can be developed also if the ordering is merely weak; see Spohn 2012: sct. 14.10).

Readers will have to evaluate this assumption for themselves. Suffice it to say that, as Geiger & Pearl (1988) show, the assumption that the variables are linearly ordered by time (and that variables causally relevant to others precede the latter) implies the following: the "causal structure" of a causal model, i.e., its directed acyclic graph, is fixed by the conditional independence relation of any probability measure satisfying the Markov and minimality conditions for this graph. In fact, this is so for not only conditional probabilistic independence, but any relation, such as conditional rank-theoretic independence, that satisfies the graphoid axioms (Spohn 1978; 1980; 1994; 2001, Dawid 1979, Studený 2005). Without this assumption, the conditional independence relation of a probability measure may satisfy the Markov and minimality conditions for different causal structures (see Hitchcock 2018 for these conditions and an example).[2]

Whether successful or not, the point now is that this account of objectification is not applicable to probability measures. Presumably, this is why Spohn (1999; 2010) ends up with objective probability as primitive. However, then we have to live with alethic modality anyway: the motivation for objectification vanishes. In fact, on the theories of chapter 5, the situation is worse for Spohn (2012) than for Skyrms (1980). On these theories, both appeal to counterfactuals through conditional degree of certainty. However, Spohn (2012) additionally appeals to counterfactuals through grade of non-conditional disbelief, conditional belief, and grade of conditional disbelief, as well as to objective probability.

[2]In other words, the information in the directed acyclic graph of a causal model goes beyond the information in its probability measure. In the same way, the information in the structural equations, if present, of a causal model goes beyond the information in its directed acyclic graph. Even this information may not be all there is to causality, as we will see in the next chapter.

On the theories of chapter 5, Spohn (2012) already appeals to counterfactuals multiply. So, there is no reason not do to so again. Indeed, Occam's razor urges us to. For now actual causation can be defined without appealing to (the direction of) time. The cognitivist desire of assigning truth conditions to claims about actual causation is achieved by interpreting the ranking function in the modal idealist's singular terms of descriptive normality. Doing so delivers the "objectivity" of actual causation Spohn (1993; 2012: ch. 15) aims at without loss of associationist flexibility in theorizing or cognitivist cost in ontological parsimony. (The frame-relativity or model-dependence of actual causation, to be discussed in the next chapter, remains. According to modal idealism, ideas are relative to a language. Frame-relativity makes this dependence explicit. It is welcome and water on the modal idealist's mills.) In addition, this renders causal relevance between events a special case of direct causation and, hence, Lewis (1973b)'s theory of actual causation an instance of Spohn (2006)'s. This in turn may be why, to this day (Kroedel 2020), counterfactual dependence, i.e., causal relevance between events, is often considered to be sufficient for actual causation.

How do we get these results? Recall that counterfactuals receive truth values relative to presuppositions or contexts (section 7.4). So, let us assume that the relevant context for direct causation of event e by event c in possible world w is the complete history of w up to right before the effect e, but excluding the cause c, $H_w^{c,e}$. Then C, which is a consequence of the causal relevance of c to e in w, if the typicality function of w is regular, implies D:

$$r_w\left(E \cap \overline{C} \cap H_w^{c,e}\right) > r_w\left(\overline{E} \cap \overline{C} \cap H_w^{c,e}\right) \text{ and } r_w\left(\overline{E} \cap C \cap H_w^{c,e}\right) > r_w\left(E \cap C \cap H_w^{c,e}\right)$$

Let us also assume that the agent whose conditional grades of disbelief deliver the correct judgments of direct causation on Spohn (2006)'s subjective theory is certain of the theory of deterministic alethic modality of possible world w, D_w, and obeys the royal rule (section 7.4). Then D implies E:

$$R\left(E \mid \overline{C} \cap H_w^{c,e}\right) > R\left(\overline{E} \mid \overline{C} \cap H_w^{c,e}\right) \text{ and } R\left(\overline{E} \mid C \cap H_w^{c,e}\right) > R\left(E \mid C \cap H_w^{c,e}\right),$$

which says that, in w, c is a necessary and sufficient cause of e.

Of course, my friendly suggestion to theorists of actual causation is to replace R and P, or similar inequalities, by the following, R* and P*, respectively:

$$\begin{aligned} r_w\left(\overline{E} \cap C \cap H_w^{c,e}\right) - r_w\left(E \cap C \cap H_w^{c,e}\right) &> r_w\left(\overline{E} \cap \overline{C} \cap H_w^{c,e}\right) - r_w\left(E \cap \overline{C} \cap H_w^{c,e}\right) \\ \log\left(\frac{ch_w\left(E \cap C \cap H_w^{c,e}\right)}{ch_w\left(\overline{E} \cap C \cap H_w^{c,e}\right)}\right) &> \log\left(\frac{ch_w\left(E \cap \overline{C} \cap H_w^{c,e}\right)}{ch_w\left(\overline{E} \cap \overline{C} \cap H_w^{c,e}\right)}\right) \end{aligned}$$

The result is a unified theory of actual causation whose probabilistic rendering exactly parallels its rank-theoretic rendering. According to it, actual causation – like its ingredients from modal idealism: descriptive normality in its singular sense and single-case chance – is an idea we humans find extremely useful in representing, navigating through, as well as manipulating and controlling reality, but that has no reality itself.

As readers will have noticed, R^* and P^* do not appeal to conditional degree of atypicality or conditional chance. This may render these inequalities difficult to evaluate. That can be remedied, though not without cost. Chapter 7 derives the properties of descriptive normality, and sketches how one may be able to do the same for chance. However, chapter 7 does not mention conditional degree of atypicality or conditional chance. Of course, these notions can be introduced. Furthermore, it can be stipulated that these conditional notions relate to their non-conditional counterparts through the difference and ratio formula, respectively. That being said, if this stipulation is one of definition, nothing has been gained: the result is definitionally equivalent to R^* and P^*.

By contrast, if this stipulation is an assumption about the relation between conditional and non-conditional degrees of atypicality and chances, it requires theories of conditional degree of atypicality and chance, as well as arguments that the resulting notions satisfy the difference and ratio formula, respectively. The former task can perhaps be achieved by formulating theories of conditional normality and chance along the lines of the conditional theories of conditional belief and degree of certainty from chapter 5. (To illustrate, here is the conditional theory of conditional chance: at w, the conditional chance of A given C equals x if, and only if, at w, the chance of C equals 1 and the chance of the material conditional $C \rightarrow A$ equals x, or, at w, the chance of C does not equal 1 and the chance of the material conditional $C \rightarrow A$ would equal x if the chance of C, but no logically stronger proposition, were equal to 1, and this were all that directly affects chances at w.) The latter task requires substantially more work.

Finally, readers may wonder what happened to the distinction between causal or interventionist counterfactuals on the one hand, as well as backtracking and spurious or acausal counterfactuals on the other hand. On the modal idealist's version of Spohn's theory, R^*, it is the background condition $H_w^{c,e}$, as well as the interpretation of the formal framework of singular variables, that guarantee that counterfactuals do not backtrack and are not spurious. Of course, this presupposes the adequacy of Spohn's theory of actual causation, modulo the interpretation of the ranking function. The next chapter will show how to characterize causal counterfactuals without this assumption; the next section will set the stage for it.

8.3 Causal counterfactuals and Arrow's theorem

It is time to address the question of what distinguishes causal or interventionist counterfactuals from others such as backtracking, as well as spurious or acausal counterfactuals. To this end, I will first discuss Lewis (1979)'s answer to this question. Then I will raise a problem for this answer that I have first learned of from Thomas Kroedel in Konstanz in July 2009, when Samir Okasha presented (a precursor of) Okasha (2011). The problem arises from an application of Arrow (1951)'s impossibility theorem and is also raised in Morreau (2010). Finally, I will sketch Kroedel & Huber (2013)'s way out of this problem. This way out depends on replacing the formal structure of a merely comparative preference relation by the formal structure of a quantitative ranking function. A different way out that is applicable to both of these formal structures is a byproduct of the next chapter.

According to Lewis (1973a), a counterfactual $\alpha \boxdot\!\!\rightarrow \gamma$ is true at a possible world w if, and only if, there is no possible world which is accessible from w and in which α is true; or there is a possible world which is accessible from w, in which α and γ are true, and which is, overall, more similar to w than any possible world in which α is true and γ is false. The conditions Lewis (1973a) imposes on overall similarity between possible worlds do not distinguish between causal counterfactuals and other counterfactuals. This is done only by the "system of weights or priorities" from Lewis (1979: 472). However, counterfactuals are said to be "notoriously vague" (Lewis 1973a: 1) and the vagueness is claimed to be partly resolved by context (Lewis 1973a: 67). According to Lewis (1979: 457), this (partial) resolution of vagueness through context determines also whether a counterfactual is (1) causal: what is held fixed is what remains of a certain causal structure after holding fixed what is not causally downstream of the antecedent; (2) backtracking: what is held fixed is all of this causal structure; or (3) spurious: what is held fixed are elements of a structure that is not causal in the narrow sense of efficient causation (Aristoteles BCE/1984), though perhaps in some wider sense, including grounding (Wilson 2018), or no elements of any structure are held fixed.

Causal counterfactuals result from the "standard resolution" of vagueness, while backtracking counterfactuals result from a different, "special resolution" (Lewis 1979: 457). For Lewis (1973b; 2000), who is a realist about causation and who aims at a reductive analysis of actual causation in *entirely* acausal terms – not just potentially causal terms other than 'actual causation' – this causal structure is objectively given and needs to be specified in entirely acausal terms. As we will see in the next chapter, other authors think of this causal structure neither as objectively given nor as amenable to reductive analysis in (entirely) acausal terms.

Before presenting Lewis (1979)'s answer to our question, I need to register a first difference to the view developed in the next chapter. On the latter view, it is not context which determines, through the resolution of vagueness or, perhaps, in some other way, whether a counterfactual is causal. Rather, there are different connectives that have some features in common – for instance, they all satisfy the logical postulates 1.-10^{+}. from section 7.4 – but differ in other respects. Whether an utterance expresses a causal counterfactual, a different counterfactual, or no counterfactual at all is not determined by context, but by the use of grammatical constructions (unless context were construed overly narrowly so that what is said on a particular occasion would itself be part of context, a view Lewis 1973a: 13 too rejects). In particular, in many contexts one can choose to express a causal counterfactual or a different one.

To illustrate, suppose whether or not Ida sleeps in, as well as whether or not she goes for a run in the morning, is directly causally determined by whether or not she has wine the night before: Ida would go for a run in the morning if, and only if, she were to have wine the night before; Ida would sleep in if, and only if, she were to have wine the night before. This is the causal structure that is held fixed in its entirety by the backtracking counterfactual BC, and to the extent possible by the causal counterfactual CC. In addition, there are three facts: Ida has wine the night before; she sleeps in; and, she goes for a run in the morning. The following two sentences express, respectively, a causal counterfactual and a backtracking counterfactual that are (or can be imagined to be) true in this situation:

CC *Even* if Ida had not slept in, she would *still* have had wine the night before, and she would *still* have gone for a run in the morning.

BC If Ida had not slept in, she *must* not have had [*that* would have been *because* she did not have] wine the night before, and, *so*, she would not have gone for a run in the morning.

The words in italics highlight the relevant grammatical constructions that indicate which connective is used and which counterfactual is expressed. In many contexts one can choose to use both connectives, as well as express both counterfactuals. It just is not possible to do both of these things in a single context. The reason is that holding fixed what is not causally downstream of the antecedent and, at the same time, not holding fixed what is not causally downstream of the antecedent is possible only by not holding fixed the causal structure itself. The latter in turn means that 'causally upstream' and 'causally downstream' do not have a fixed meaning anymore.

With this out of the way, let us turn to Lewis (1979)'s characterization of causal counterfactuals. In response to the future similarity objection by Fine (1975) and others (see Lewis 1979: 467) discussed in section 7.4, Lewis (1979: 472) argues that overall similarity between possible worlds,

> taken under the standard resolution of vagueness, must be governed by the following system of weights or priorities.
>
> (1) It is of the first importance to avoid big, widespread, diverse violations of law.
>
> (2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
>
> (3) It is of the third importance to avoid even small, localized, simple violations of law.
>
> (4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

The "law[s]" are the causal structure that is held fixed by both backtracking and, to the extent possible, causal counterfactuals. The "particular fact[s]" are what is causally upstream and downstream of the antecedent. The asymmetry between "big, widespread, diverse violations of law" in (1) and "small, localized, simple violations of law" in (3) – between "big miracles" and "small miracles" (Lewis 1986e: 55f) – guarantees that the particular facts in (2) (and (4)) that are held fixed by causal, but not backtracking, counterfactuals are not causally downstream.

As mentioned, Lewis (1973b; 2000) is not only a realist about causation, but also aims at a reductive analysis of actual causation in acausal terms. Therefore, he needs to analyze the ingredients of this system – that is, laws of nature and particular facts, as well as size of miracles (and similarity of particular fact) – in acausal terms, or else take them as acausal primitives. Particular facts are taken as acausal primitive, and we can grant that they are objectively given in a sense acceptable to a realist about causation. For the sake of argument let us grant the same for size of miracles (and similarity of particular fact).

Laws of nature are supposed to be analyzed in acausal terms by the best system analysis (Lewis 1973a: 73):

> a contingent generalization is *a law of nature* if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength.

We can grant that the best system analysis succeeds in analyzing laws of nature in acausal terms. The more pressing question is whether it also renders laws of nature objectively given in a sense acceptable to a realist about causation.

One possibility for the best system analysis to render laws of nature objectively given is for there to be measures of objective strength and objective simplicity that are combined in an objectively correct way. This possibility is not very likely for several reasons one of which is the following. The measures of strength that come to mind (e.g., Bar-Hillel 1952; 1955, Carnap & Bar-Hillel 1952, Bar-Hillel & Carnap 1953, Hintikka & Pietarinen 1966, and Levi 1967; see also Huber 2008: 97ff[3]) are all captured in terms of Carnap (1945)'s probability$_1$. According to Lewis (1980: 263), the latter is subjective, not objective.

Another possibility is for there to be measures of partially subjective strength and partially subjective simplicity that are combined in a way that is objectively correct and, in addition, results in an objective notion of simplicity-cum-strength. Among others, this means that this notion is insensitive to the subjective nature of its input. Perhaps partially subjective simplicity is inversely proportional to partially subjective strength, and the two are combined in a way that is invariant to shifts in its input: an increase / decrease in partially subjective strength amounts to a decrease / increase in partially subjective simplicity such that the combined degree of overall simplicity-cum-strength remains unaffected.

This possibility, too, is not very likely for several reasons one of which is the following. For it to actualize, one is in need of not only a quantitative notion of simplicity, but also one that stands in the right relation to strength. However, on at least one conception of simplicity (Kelly 2007, Kelly et al. 2016), it is *not* the case that "the virtues of simplicity and strength tend to conflict" (Lewis 1973a: 73). Consider the standard example of simplicity, curve-fitting: a linear curve is simpler than a quadratic curve. On the present conception, this is so because the former can be falsified by only three data points, while at least four data points are required to falsify the latter. However, to the extent that simplicity goes hand in hand with falsifiability (Popper 1935), the virtues of simplicity and strength tend to agree rather than conflict. (Simplicity and strength tend to agree also on Sober 1975's conception of simplicity, although he has since changed his view; see Sober 1988; 1990; 2015.)

[3]I would like to use this opportunity to correct a mistake in Huber (2008: 101) pointed out to me by Christopher Pariso (personal communication in 2015). $\inf(H)$ equals $-\log_2 \Pr(H \mid E \wedge B)$, not $-\log_2 \Pr(\neg H \mid E \wedge B)$. So, $E(\inf(H))$ is positive / negative if, and only if, $\Pr(H \mid E \wedge B)$ is smaller / greater than $\Pr(\neg H \mid E \wedge B)$, not the other way round. Furthermore, inf recommends neither maximizing nor minimizing probability, but peaks if $\Pr(H \mid E \wedge B) \approx 0.16$.

The defender of the best system analysis is, of course, free to adopt a different conception of simplicity. However, in this case the question arises whether "the virtue[...] of simplicity" is virtuous at all. (Sober 1988; 1990; 2015 suggests that simplicity reduces to different other virtues in different local contexts, and possibly to no virtue in some local contexts. The answers mentioned in Sprenger & Hartmann 2019: ch. 11, I think it is fair to say, fall short of the ones by Kelly 2007 and Kelly et al. 2016.)

The latter problem persists if simplicity or strength is a merely comparative concept. Furthermore, combining merely comparative concepts faces additional difficulties of its own. This is so if at least one of these comparative concepts is not connected so that not any two pairs of, in our case, true deductive systems can be compared with respect to this concept (Feyerabend 1962; Kuhn 1962); if at least one of these concepts is connected, but not transitive (Ramsey 1926, Davidson et al. 1955, although the concept is a different one); and if at least one of these concepts is not reflexive: what should it mean that some true deductive system is not at least as simple or strong as itself? This is so also if all of these comparative concepts are reflexive, transitive, and connected, as we will see shortly.

I conclude that Lewis (1979: 472)'s "system of weights or priorities" has the best chance of providing the basis for a reductive analysis of actual causation in acausal terms that satisfies a realist about causation if laws of nature, too, are taken as acausal primitive (or are analyzed in acausal terms by some alternative to the best system analysis). This in turn brings to the fore the following question: why would a realist distinguish between different kinds of fact – between contingent laws of nature on the one hand and matters of particular fact on the other hand – in the first place? The question arises also for expressivists and idealists, but is less problematic for them: the former can perhaps locate the source of this distinction in various subjective elements of mental states, and the latter can locate it in elements different subjective mental states have in common. However, for a realist the introduction of this distinction is yet another violation of Occam's razor. I can understand why a realist would *want* reality to be lawful, but that helps only with wishful thinking, not with philosophy. I can understand also why beings evolve to have systems of representation that render the most useful representation – in the sense of section 7.3: the actual world – and maybe others lawful. However, that is of help only to an idealist (and, perhaps, an expressivist), not also a realist. I remain puzzled. Of course, just because I am puzzled does not mean that there is anything problematic (for anyone other than me). Unfortunately, as Kroedel & Huber (2013) show, there is another issue with Lewis (1979)'s characterization of causal counterfactuals that is a problem not only for me.

The general picture of overall similarity that Lewis (1973a; 1979) appears to have is as follows: whether possible world u is, overall, more similar to possible world w than possible world v is to w is determined by various individual aspects with respect to which u is more similar to w than, as similar to w as, or less similar to w than v is to w. Lewis (1979: 472)'s "system of weights or priorities" mentions four such individual aspects of similarity: whether u is more similar to w than, as similar to w as, or less similar to w than v is to w (1) with respect to "avoid[ing] big, widespread, diverse violations of [w-]law;" (2) with respect to "maximiz[ing] the [w-]spatio-temporal region throughout which perfect match of particular [w-]fact prevails;"[4] (3) with respect to "avoid[ing] small, localized, simple violations of [w-]law;" and (4) with respect to "secur[ing] approximate similarity of particular [w-]fact."

Lewis (1973a) stipulates that the relation of overall similarity between possible worlds is, among others, reflexive, transitive, and connected, as well as merely comparative. With Kroedel & Huber (2013: 455), I assume the same to be true of the four relations of similarity with respect to an individual aspect. My reasons are the following. A merely comparative relation (of similarity with respect to an individual aspect) that is not connected – or connected, but not transitive – faces other issues, as mentioned in the previous paragraphs (admittedly for different concepts, but these issues carry over, or so I assume). A merely comparative relation that fails to be reflexive does not make sense: what should it mean that some possible world u is not at least as similar to possible world w in avoiding big w-miracles, maximizing perfect match of particular w-fact, avoiding small w-miracles, or securing approximate similarity of particular w-fact as u is to w? (As an aside, note that a relation of overall similarity between possible worlds can be reflexive without being weakly centered – see section 7.4 – as well as weakly centered without being reflexive.)

Therefore, the determination of overall similarity between possible worlds by individual aspects of similarity between possible worlds has the same structure as the social choice problem (Arrow 1951): the question of how to aggregate the individual preferences of various members of society over a number of alternatives into an overall preference of society itself over these alternatives. As Kroedel & Huber (2013) argue, it is also subject to the same criteria.

[4] I relativize 'spatio-temporal' to a possible world because the meaning of 'space-time' may depend on contingent physical theory: arguably, space and time are different things in Newton (1687) and Einstein (1905). If so, there may be still further problems for (1-4) and Lewis (1986a)' modal realism – especially if one adopts the best system analysis and the best systems for some possible worlds do not mention space or time.

To state these, let us focus on overall similarity, as well as individual aspects of similarity, *to the actual world*. An *overall similarity ordering* is a relation among possible worlds that specifies, for all possible worlds u and v, whether, overall, u is more similar to the actual world than, as similar to the actual world as, or less similar to the actual world than v is to the actual world. An *aspectual similarity ordering* is a relation among possible worlds that specifies, for all possible worlds u and v, whether, with respect to an individual aspect of similarity, u is more similar to the actual world than, as similar to the actual world as, or less similar to the actual world than v is to the actual world. A *profile* is a sequence of n aspectual similarity orderings, for a fixed, finite natural number $n \geq 2$ such as 4.

With this terminology in hand, our problem can be stated as follows. We are looking for a function f from the set of profiles into the set of overall similarity relations that satisfies the following four conditions.

P (Weak Pareto Principle) If possible world u is more similar to the actual world than possible world v is to the actual world according to all individual aspects of similarity, then u is, overall, more similar to the actual world than v is to the actual world.

I (Independence of Irrelevant Alternatives) If two profiles do not differ with respect to possible worlds u and v, then the two overall similarity orderings determined for them by f do not differ with respect to u and v either.

U (Unrestricted Domain) The domain of the function f includes all profiles that are mathematically possible.

D (Non-Dictatorship) There is no individual aspect of similarity such that for all possible worlds u and v: if u is more similar to the actual world according to this aspect than v is to the actual world, then u is, overall, more similar to the actual world than v is to the actual world.

According to Arrow (1951)'s impossibility theorem (see also Gaertner 2009: 19-21 and Morreau 2010), if there are at least three possible worlds, then there is no function f satisfying all of P, I, U, and D. This means that Lewis (1979: 472)'s "system of weights or priorities" cannot be a system of weights. As Kroedel & Huber (2013: 456) and others (*ibid.* fn. 6) argue, on pain of being incorrect, it also cannot be a system of priorities – that is, a lexicographic order according to which similarity to the actual world with respect to perfect match of particular actual-world-fact matters only as a tiebreaker for possible worlds that are equally similar to the actual world with respect to avoiding big actual-world miracles; etc.

What is the similarity theorist to do? In social choice theory, Sen (1970)'s suggestion for solving the social choice problem is to equip the various members of society with a numerical utility function on the set of alternatives rather than a merely comparative preference relation over the latter. The additional information provided by these functions makes it possible to aggregate them into a numerical utility function on the set of alternatives for society itself, and to do so in a way that satisfies four conditions analogous to P, I, U and, D. For instance, one can solve the social choice problem in this way by simply adding up the numerical utility values assigned by the various members of society to each alternative. Of course, this raises other questions such as whether the numerical utility values different members of society assign to an alternative can be compared to each other – that is, whether interpersonal utility comparisons are possible.

Kroedel & Huber (2013) offer the similarity theorist a way out that parallels Sen (1970)'s suggestion. Instead of merely comparative relations of similarity with respect to individual aspects, they suggest to consider numerical functions of aspectual similarity on the set of possible worlds. These functions of numerical similarity can be aggregated into a numerical function of overall similarity on the set of possible worlds, as well as a merely comparative relation of overall similarity between possible worlds that satisfies all of Lewis (1973a)'s formal constraints (plus the limit assumption; see section 7.4). Furthermore, this can be done such that the following four conditions are satisfied.

P* (Weak Pareto Principle) If possible world u is more similar to the actual world than possible world v is to the actual world according to all individual aspects of *numerical* similarity, then u is, overall, more similar to the actual world than v is to the actual world.

I* (Independence of Irrelevant Alternatives) If two *numerical* profiles do not differ with respect to possible worlds u and v, then the two overall similarity orderings determined for them by f do not differ with respect to u and v either.

U* (Unrestricted Domain) The domain of the function f includes all *numerical* profiles that are mathematically possible.

D* (Non-Dictatorship) There is no individual aspect of *numerical* similarity such that for all possible worlds u and v: if u is more similar to the actual world according to this aspect than v is to the actual world, then u is, overall, more similar to the actual world than v is to the actual world.

On top of this, Kroedel & Huber (2013: 462ff) argue that this aggregation does not run into issues of incommensurability parallel to those in interpersonal utility comparisons: where different aspects of similarity are difficult to compare, this is due to vagueness, not incommensurability.

Should the similarity theorist adopt this way out? No, she should revise her theory. As we will see in the next chapter, it is not the size of miracles that matters, but when in the causal history they occur. Furthermore, while the relative importance of miracles will be specified in terms of a lexicographic order, it is perfect match of *certain* particular fact and avoidance of *certain* miracles that matter, and do so equally. Other particular fact and avoidance of other miracles does not matter less or infinitely less, but not at all. Finally, once all of this is fixed, causal counterfactuals can be characterized also in merely comparative terms.

To illustrate, consider again the causal counterfactual CC and the backtracking counterfactual BC whose antecedents say that Ida does not sleep in. Let w be a possible world in which Ida has wine the night before, she sleeps in, she goes for a run in the morning, and whose first and second "law" is, respectively, that Ida would sleep in if, and only if, she were to have wine the night before and that Ida would go for a run in the morning, if and only if, she were to have wine the night before. These "laws" specify the relevant causal structure of w. BC holds fixed all of this causal structure. CC holds fixed what remains of it after holding fixed what is not causally downstream of the antecedent. According to Lewis (1973a; 1979), CC is (non-vacuously) true at w if, and only if, there is a possible world that is accessible from w and in which Ida has wine the night before, she does not sleep in, and she goes for a run in the morning; and, this possible world is, overall, more similar to w than every possible world in which Ida does not sleep in and, in addition, she does not have wine night before or she does not go for a run in the morning. Any such possible world holds fixed what is not causally downstream of the antecedent – namely, that Ida has wine the night before – even if this means violating the causal structure of w, namely, the first "w-law." Only after holding fixed what is not causally downstream of the antecedent (which, in this example, coincides with what is causally upstream of the antecedent), does it matter to avoid violating what remains of the causal structure of w, namely, the second "w-law." Of course, Lewis (1973a; 1979) could respond that our "laws" are not laws of nature or that they are, but that their violations constitute small rather than big miracles. But then we are not only back at the questions of what a law of nature is, and how to measure the size of miracles. We also face a new one: why is it enough to specify three facts and a simple causal structure to determine that CC rather than BC comes out as true causal counterfactual?

Chapter 9

Causality and Counterfactuals

In this chapter I will first present the formal framework of extended causal models. These involve, among others, structural equations which represent a certain causal structure, as well as an assignment of normality or typicality. Then I will slightly generalize this framework to facilitate relating it to a different, more parsimonious framework for counterfactuals and default conditionals: typicality models. Next I will formulate four constraints on extended causal models. One of these relates structural equations and normality and characterizes backtracking counterfactuals. Another relates structural equations, normality, and actuality and characterizes causal counterfactuals. Finally, I will generalize typicality models to causality models which go beyond extended causal models in several respects and render interventions definable and my central constraint provable. I rely on Huber (2013).

9.1 Causal models

The most promising framework for theorizing about causality, representing causal structure, and analyzing actual causation seems to be the causal models approach (Spirtes et al. 1993/2000 and Pearl 2000/2009; see also Halpern & Pearl 2005a; b). The following definitions follow Halpern (2008).

$\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$ is a *causal model* if, and only if, $\mathcal{S}$ is a signature and $\mathcal{F} = \{F_1, \ldots, F_n\}$ represents a set of n structural equations, for a finite natural number n. $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ is a *signature* if, and only if, $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V} = \{V_1, \ldots, V_n\}$ is a set of n endogenous variables that is disjoint from $\mathcal{U}$, and $R : \mathcal{U} \cup \mathcal{V} \rightarrow \mathcal{R}$ assigns to each exogenous or endogenous variable X in $\mathcal{U} \cup \mathcal{V}$ its *range* (not co-domain) $R(X) \subseteq \mathcal{R}$.

$\mathcal{F} = \{F_1, \ldots, F_n\}$ *represents a set of n structural equations* if, and only if, for each natural number i, $1 \leq i \leq n$: F_i is a function from the Cartesian product $\mathcal{W}_i = \times_{X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_i\}} R(X)$ of the ranges of all exogenous and endogenous variables other than V_i into the range $R(V_i)$ of the endogenous variable V_i. The *set of possible worlds* of the causal model $\mathcal{M}$ is defined as the Cartesian product $\mathcal{W} = \times_{X \in \mathcal{U} \cup \mathcal{V}} R(X)$ of the ranges of all exogenous and endogenous variables.

A causal model $\mathcal{M}$ is *acyclic* if, and only if, it is not the case that there are m endogenous variables $V_{i1}, \ldots, V_{im}$ in $\mathcal{V}$, for some natural number m, $2 \leq m \leq n$, such that the value of $F_{i(j+1)}$ depends on $R\left(V_{ij}\right)$ for $j = 1, \ldots, m-1$, and the value of F_{i1} depends on $R(V_{im})$. Importantly, dependence is just ordinary functional dependence: F_i depends on $R\left(V_j\right)$ if, and only if, there are arguments $\vec{w}_i$ and $\vec{w}_i{}'$ in the domain $\mathcal{W}_i = \times_{X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_i\}} R(X)$ of F_i that differ only in the value from $R\left(V_j\right)$ such that their values under F_i differ, $F_i(\vec{w}_i) \neq F_i\left(\vec{w}_i{}'\right)$.

Let $Pa(V_i)$ be the set of variables X in $\mathcal{U} \cup \mathcal{V}$ such that F_i depends on $R(X)$. The elements of $Pa(V_i)$ are the *parents* of the endogenous variable V_i, that is, the set of variables that are *directly causally relevant* to V_i. Let $An(V_i)$ be the ancestral, or transitive closure, of $Pa(V_i)$, which is defined recursively as follows: $Pa(V_i) \subseteq An(V_i)$; if $V \in An(V_i)$, then $Pa(V) \subseteq An(V_i)$; and, nothing else is in $An(V_i)$. The elements of $An(V_i)$ are the *ancestors* of the endogenous variable V_i. A *context* is a specification of the values of all exogenous variables. It can be represented by a vector $\vec{u}$ in the Cartesian product $R(\mathcal{U}) = \times_{U \in \mathcal{U}} R(U)$ of the ranges of all exogenous variables. A basic fact about causal models is that every acyclic causal model has a unique solution for any context. An acyclic causal model can be pictured by a directed acyclic graph whose nodes are the exogenous and endogenous variables in $\mathcal{U} \cup \mathcal{V}$ and whose arrows point into each endogenous variable V_i from all of the latter's parents in $Pa(V_i)$. In general, the directed acyclic graph contains less information than the acyclic causal model it pictures, as different structural equations can give rise to the same nodes and arrows. Sometimes a picture says less than a few words.

The signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ provides the language of the causal model $\mathcal{M}$. It has more structure than the set of possible worlds W of a model $\langle W, (\$_w)_{w \in W} \rangle$ in the traditional possible worlds semantics, where, for any possible world w in W, $\$_w$ is an accessibility relation over W or, say, a system of spheres, probability measure, ranking function, or something of this kind. The reason is that there is a distinction between exogenous and endogenous variables. That said, in causal models this distinction can be recovered from the structural equations (by defining as exogenous those variables for which there is no structural equation).

What is just as important is how one understands these variables. I understand them as singular, not generic variables (see section 8.2).

Philosophers such as Woodward (2003), following Spirtes et al. (1993/2000) and Pearl (2000/2009), are mainly interested in causal relevance between generic properties rather than actual causation between events (or similar entities such as facts, aspects, and tropes). This means they understand the variables in the generic sense they are usually understood in the sciences. Generic variables assign values to the individuals of a population from which one can draw samples. For instance, the population may be the set of people at a certain age and in a certain region, and the generic variable may assign value k to an individual in this population if this individual consumes k mg ibuprofen.

On this generic understanding of the variables it may well be possible to test *generic* causal counterfactuals by "carry[ing] out the interventions described in the[...] antecedents and then check[ing] to see whether certain correlations hold" (Woodward 2003: 72-73). For instance, it may well be possible to test in this way whether the generic property of administering ibuprofen is causally relevant to the generic property of relief of pain by carrying out the intervention of administering a certain amount of ibuprofen to some people in the population and then checking to see if pain is relieved in them. Unfortunately, the semantics of generic claims – whether they involve causal counterfactuals, default rules (chapter 1 and section 7.2), or neither – is notoriously unclear (Leslie & Lerner 2016). Specifically, to the best of my beliefs, there is no semantics for generic causal counterfactuals. So, it is not clear what exactly the claims tested in this way mean.

In contrast to this, a precise semantics can be stated for non-generic causal counterfactuals in terms of possible worlds (or related semantic items such as Fine 2012a's possible states) in several ways. One of these is Galles & Pearl (1998)'s structural equations semantics outlined in section 9.3 (see also Halpern 2000; 2013, Hiddleston 2005a, Briggs 2012, Zhang 2013, and Zhang et al. 2013). Crucially, though, this requires the variables to be understood in a *singular* sense. Singular variables assign values to possible worlds that are mutually exclusive so that one cannot draw samples from their collection. Hence, a variable now assigns value k to a possible world if, in this possible world, a specific person consumes k mg ibuprofen on a specific occasion, say, Ida in the morning of her 30th birthday. With this singular understanding of the variables we can then construct the set of possible worlds by forming the Cartesian product of the ranges of all variables in the way we have done in the preceding paragraphs. This in turn means that both the antecedent and consequent of a causal counterfactual express a proposition or set of possible worlds (or states), as does the causal counterfactual itself.

Figure 9.1: Ida sleeps in on her 30th birthday

The good news is that interpreting the variables in a singular sense to state a precise semantics for causal and backtracking counterfactuals does not mean that we lose the ability to test these claims. As we will see in the next chapter, both backtracking and causal counterfactuals can be tested "empirically." In a nutshell, a backtracking counterfactual is tested by passively "observing" that an antecedent obtains (so that one can assume that it occurs by being caused by its direct causes). An interventionist counterfactual – that is, one for which there is a structural equations semantics – is tested by actively carrying out a "hard intervention" to bring about an antecedent (so that one's intervention is the sole direct cause of the occurrence of the antecedent). Finally, our more general causal counterfactual of which an interventionist counterfactual is a special case is tested by actively carrying out a "possibly soft intervention" (so that one's intervention is among the direct causes of the occurrence of the antecedent, possibly besides others). Before studying in the next chapter under what conditions these claims can be tested in what sense, let us use this chapter to state their precise semantics.

There is another reason why the interpretation of the variables matters. In many cases, it affects whether a causal model is acyclic (see also Kistler 2013). To illustrate, consider a causal model with the following singular variables: W_1 takes on value 1 if Ida has wine the night before her 30th birthday, and 0 otherwise; SL takes on value 1 if Ida sleeps in on her 30th birthday, and 0 otherwise; LW takes on value 1 if Ida is late for work on her 30th birthday, and 0 otherwise; ST takes on value 1 if Ida is stressed on her 30th birthday, and 0 otherwise; and W_2 takes on value 1 if Ida has wine in the evening of her 30th birthday, and 0 otherwise. There are four structural equations represented by F_{SL}, F_{LW}, F_{ST}, and F_{W_2}, respectively, which say that W_1, SL, LW, and ST, respectively, directly causally determines the value of SL, LW, ST, and W_2, respectively: the latter variable takes on value 1 (0) if the former variable takes on value 1 (0). For instance, F_{SL} says that Ida would sleep in on her 30th birthday if, and only if, she were to have wine the night before her 30th birthday, and similarly for the other three structural equations. W_1 is the sole exogenous variable. Since the causal model is acyclic, it can be pictured by a directed acyclic graph (figure 9.1) whose nodes are the exogenous and endogenous variables and whose arrows point into each endogenous variables from all of the latter's parents.

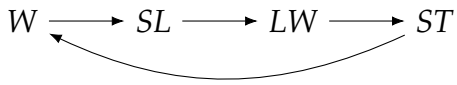

Figure 9.2: Ida sleeps in

Now change the interpretation of the variables from singular to generic and let the population be the days of Ida's life: W assigns value 1 to a day of Ida's life if Ida has wine in the evening of this day, and 0 otherwise; similarly for SL, LW, and ST. In this case it is not clear if the causal model should still be acyclic. Instead, it may be that now an adequate causal model is a cyclic "feedback system" pictured by a directed cyclic graph (figure 9.2). In this graph, not only the nodes, but also the arrows represent different things that cannot anymore be stated in terms of causal counterfactuals. Instead, the arrows and underlying structural equations now encode generic causal counterfactuals and it is up to the proponent of generic variables to tell us what exactly these mean.

It is worth noting that these concerns are less pressing for probabilistic causal models without structural equations. These models still picture the variables with the nodes of a directed acyclic graph. However, they assume the arrows of the directed acyclic graph to be given independently rather than being derived from a set of structural equations. In addition, there is a probability measure which has to satisfy several constraints relative to the directed acyclic graph (see section 8.2). In general, the directed acyclic graph of a probabilistic causal model without structural equations contains more information than the conditional independence relation of its probability measure. But as shown by Geiger & Pearl (1988), and of particular relevance for attempts to reduce causality to probability, if the variables are linearly ordered, the directed acyclic graph can be read off the conditional independence relation of the probability measure (again, see section 8.2).

Here too the arrows mean different things depending on whether the variables are interpreted in a singular or generic sense. The difference here is that, unlike causal counterfactuals, probability is meaningful for both interpretations of the variables. For singular variables one can adopt a chance or degree of certainty conception of probability. For generic variables one can adopt a relative frequency conception. Against this background the results of this chapter acquire additional significance: we will see that structural equations can be reproduced in typicality models. This opens the door for a generic interpretation of the variables of a causal model. The reason is that, like probability, typicality too can be conceived in a singular, as well as a generic sense (see chapter 1 and section 7.2).

In concluding this section, suppose that we have a probabilistic acyclic causal model with structural equations. The structural equations determine a directed acyclic graph as explained previously. Let us assume that the probability measure satisfies the constraints mentioned previously relative to this graph. In this case the structural equations contain, in general, more information than the directed acyclic graph which in turn contains, in general, more information than the conditional independence relation of the probability measure. So, it is the structural equations that we need to, and will, capture. In fact, we will go beyond them.

9.2 Extended causal models

To motivate the introduction of normality, let us return to the analysis of actual causation (section 8.2). Pearl (2000/2009: ch. 10), Hitchcock (2001), Woodward (2003: ch. 2), and Halpern & Pearl (2005a) provide increasingly sophisticated definitions of actual causation in terms of acyclic causal models. These provide reductive analyses, as they are stated in terms other than 'actual causation.' They just are not stated in entirely acausal terms, as attempted by Lewis (1973b; 2000) (see section 8.3). As Hiddleston (2005b) shows, there are acyclic causal models where the "intuitively correct" causal judgments differ, even though the causal models differ only in the meaning of the variables (see section 9.4 for an example). Readers may be reminded of Goodman (1954/1983)'s new riddle of induction: the hypothesis that all emeralds are green differs only in meaning from the hypothesis that all emeralds are grue, not also its purely logical relation to the information about the color of emeralds observed so far. As Halpern (2008) puts it: "there must be more to causality than just the structural equations."

To solve this problem, Hitchcock (2007) and Hall (2007) (see Hitchcock 2009 for a reply) distinguish between normal, typical, or default values and abnormal, atypical, or deviant values of a variable. Arguably, this development is anticipated by Hitchcock (2001: 290)'s concept of the "redundancy range" of the values of the variables and Halpern & Pearl (2005: 869f)'s concept of an "allowable setting" of the endogenous variables in their "more refined" definition. In Halpern (2008) and Halpern & Hitchcock (2010) normality is represented by a ranking function. In Halpern & Hitchcock (2013; 2015) and Halpern (2016) it is represented by a comparative pre-ordering relation among possible worlds that is reflexive and transitive, but possibly not connected (section 3.1). I will first state the definitions of extended causal models and actual causation, before commenting on the formal representation of normality and its philosophical interpretation in section 9.5.

$\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, \varrho \rangle$ is an *extended* (acyclic) causal model if, and only if, $\langle \mathcal{S}, \mathcal{F} \rangle$ is a(n) (acyclic) causal model and ϱ is a ranking function on the power-set of $\mathcal{W}$. To allow (but not require) that what is normal varies from context to context, I will index the ranking function ϱ to the set of contexts. Thus, extended (acyclic) causal models really are of the form $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$, where $R(\mathcal{U}) = \times_{U \in \mathcal{U}} R(U)$ is the set of all contexts or specifications of the values of all exogenous variables.

The relata of actual causation are assumed to be representable by a variable X's taking on a specific value x from its range $R(X)$, that is, an atomic sentence of the form $X = x$, as well as the Boolean combinations that can be formed from these atomic sentences by finitely many applications of negation $\neg$, conjunction $\wedge$, and disjunction $\vee$. The variables must be endogenous. Sentences of the form $X \in S$, for a subset S of $R(X)$ with more (or less) than one element are not allowed. Anything that is representable by a Boolean combination ϕ of atomic sentences can be an effect. An actual cause must be representable by a finite conjunction $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$ of one or more atomic sentences with distinct variables. This restriction seems to have as its reason not some thesis about actual causation, but the fact that the interventionist counterfactuals ($\boxright$) employed in the definition of actual causation are not defined for antecedents of a different form. Some of these restrictions can be lifted (Briggs 2012 and Halpern 2013): one can allow for arbitrary Boolean combinations of atomic sentences as antecedents, interventionist counterfactuals in the consequent, as well as Boolean combinations of these sentences. I am not aware of suggestions allowing for counterfactuals in the antecedent. The definition of actual causation then runs as follows (Halpern & Hitchcock 2010: sct. 3 and Halpern 2016: ch. 3). $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$, or simply $\vec{X} = \vec{x}$, is an *actual cause* of ϕ in the extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ in context $\vec{u}$ if, and only if:

1. $\vec{X} = \vec{x}$ and ϕ are true in $\mathcal{M}$ in $\vec{u}$;
2. there is a partition $\left\{\vec{Z}, \vec{W}\right\}$ of the endogenous variables $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$, and there are vectors of values $\vec{x'}$ and $\vec{w}$ of $\vec{X}$ and $\vec{W}$, respectively, with $\varrho_{\vec{u}}\left(\vec{X} = \vec{x'} \wedge \vec{W} = \vec{w}\right) \leq \varrho_{\vec{u}}(w_{\vec{u}})$ such that: if $\vec{Z} = \vec{z^*}$ is true in $\mathcal{M}$ in $\vec{u}$, then
   (a) $\vec{X} = \vec{x'} \wedge \vec{W} = \vec{w} \boxright \neg\phi$ is true in $\mathcal{M}$ in $\vec{u}$, and
   (b) for all $\vec{W^-} \subseteq \vec{W}$ and all $\vec{Z^-} \subseteq \vec{Z}$: $\vec{X} = \vec{x} \wedge \vec{W^-} = \vec{w} \wedge \vec{Z^-} = \vec{z^*} \boxright \phi$ is true in $\mathcal{M}$ in $\vec{u}$; and
3. there is no proper subset $\vec{X^-}$ of $\vec{X}$ such that 1. and 2. hold for $\vec{X^-}$.

To digest this definition, recall Lewis (1973b)'s definition of causal relevance of event c for event e from section 8.2: both c and e occur; the occurrence of c is counterfactually necessary for the occurrence of e in the sense that e would not have occurred if c had not occurred; and, the occurrence of c is counterfactually sufficient for the occurrence of e in the sense that e would have occurred if c had occurred. The definitions of actual causation considered in this section can be considered variants of the definition of causal relevance between events that differ from the original, among others, in the following two respects. First, they employ counterfactuals with more specific antecedents to spell out more refined concepts of counterfactual necessity and sufficiency. Second, they drop the idea that actual causation is the transitive closure of causal relevance between events and, hence, transitive itself (for the transitivity of actual causation see Hall 2000). Arguably, these two developments start with Pearl (2000/2009: ch. 10.3)'s "causal beam." Hitchcock (2001) then considers "explicitly nonforetracking counterfactuals" and "active causal routes," while Halpern & Pearl (2005a) go on to generalize causal beams to "active causal processes." Essential to some of these notions are Spirtes et al. (1993/2000)'s paths in a directed graph. These are sequences of mutually distinct arrows such that any two adjacent arrows have a node in common. A path is directed if, and only if, all arrows point in the same direction. In addition, momentarily ignore the highlighted part so that we are dealing with Halpern & Pearl (2005a: 852f)'s "preliminary" (Halpern 2016: ch. 2's "updated") definition.

As in the definition of causal relevance between events, the first clause tells us that actual causes and effects occur. Often, but not always (Halpern 2016: sct. 2.9) $\vec{Z}$ contains variables that are on a path from one of the actual cause variables $X_1, \ldots, X_k$ to one of the effect variables occurring in ϕ. The variables in $\vec{W}$ are "bystanders." The second clause tells us that there is *some* collection of bystanders and *some* setting $\vec{w}$ of them such that this setting is witness to a weaker form of counterfactual necessity and a stronger form of counterfactual sufficiency of the actual cause for the effect. Counterfactual necessity is weakened, as one has to specify concrete alternative values $x'_1, \ldots, x'_k$ for all actual cause variables, as well as because all bystanders have to be set to their witness values. Counterfactual sufficiency is strengthened, as the actual cause has to be counterfactually sufficient for the effect as long as the remaining variables in $\vec{Z}$ are not explicitly set to non-actual values and as long as no bystander is explicitly set to a non-witness value. The third clause is in part necessitated by the weakening of counterfactual necessity: it is a minimality condition to the effect that actual causes do not contain any irrelevant conjuncts.

My paraphrase of the second clause may be somewhat misleading. It is not the case that there has to be some setting of some bystanders that is witness to the counterfactual necessity and sufficiency of the actual cause for the effect – that is, the counterfactual dependence of the effect on the actual cause. The conjunction of the concrete alternative values for all actual cause variables with the setting of all bystanders to their witness values must be counterfactually sufficient for the non-occurrence of the effect. Furthermore, each conjunction of the actual cause with any combination of actual values for the remaining variables in $\vec{Z}$ and witness values for the bystanders must be counterfactually sufficient for the effect.

I mention this, as it is tempting to think of the definitions of actual causation considered in this section as variants of the definition of causal relevance between events that differ from the original in the following way: rather than requiring the effect to be counterfactually dependent on the actual cause, they require the effect to be *conditionally* counterfactually dependent on the actual cause given some condition. (Some passages in Hitchcock 2001 suggest such a reading; for instance, Hitchcock 2001: 289; emphasis in the original: "*given* that Billy's rock did not hit the bottle, if Suzy had not thrown, the bottle would have remained intact throughout the incident.") This would not be correct. One reason is that, unlike numerical probability measures and ranking functions, as well as other functions that may be non-numerical (Halpern 2003/2017: 99ff), counterfactuals with a similarity semantics, a structural equations semantics, or Briggs (2012)' possible states semantics do not have an operation of conditionalization. Therefore, one has to resort to counterfactuals with more specific antecedents.

So, with this in mind, the idea behind Halpern & Pearl (2005a)'s definition is, crudely, that there be *some* witness to the counterfactual dependence of the effect on the actual cause. Halpern & Hitchcock (2010)'s definition, which includes the highlighted part, then requires not just that there be some witness, but that there be a witness that is *normal* (even) in the presence of alternative cause values that are *normal*, too: the conjunction of the concrete alternative values for all actual cause variables with the setting of all bystanders to their witness values must not be less normal than the conjunction of the actual values of all variables. That is, the effect must fail to occur in a possible world that is not less normal than the actual possible world. The counterfactual in part (a) of the second clause must be "relevant" (Fazelpour 2021). (Halpern 2016: ch. 2's "modified" definition, which is in the spirit of Hitchcock 2001: 286, requires that there be some bystanders such that their *actual* values are a witness. Halpern 2016: ch. 3's preferred definition additionally requires that the alternative cause values are *normal*.)

Unlike Lewis (1973b; 2000)'s, the definitions of actual causation considered in this section render actual causation relative to a model. Hitchcock (2001; 2007), Halpern & Pearl (2005a), and Halpern & Hitchcock (2010) welcome this, as do other scientists (Heckman 2005) and philosophers (Menzies 2004). Beckers & Vennekens (2017; 2018), Andreas & Günther (2021a; b; 2023), Beckers (2021a), and Fischer (2023) do not mind. Spohn (2006; 2012: sct. 14.9), Hall (2007), and Weslake (2023) do. Halpern (2016: ch. 4) studies under what conditions various changes to a model do not make a difference (see also Beckers 2021b).[1]

The present approach embraces not only modal idealism (section 7.3), but also model idealism: like any system of representation, a model is a mind-dependent construct or idea that is not objectively right or wrong, but more or less useful for various purposes. A causal claim is not true simpliciter, but true in a possible world in a model (true in a model, if the specification of the values of all variables is part of the model). The actual world of a model is the specification of the values of all variables that is most useful for the purposes of representation and – if the model contains causal aspects – manipulation and control (Woodward 2021).

9.3 Interventions

To understand Halpern & Hitchcock (2010)'s definition we still need to state the structural equations semantics for the interventionist counterfactuals of the form $\vec{X} = \vec{x} \boxright \phi$ in an extended acyclic causal model $\mathcal{M}$ in a context $\vec{u}$. It is these counterfactuals that are my target, not a definition of actual causation, whether it serves as analysis or explication (Huber 2018: sct. 4.1). To keep things readable, I will continue to ignore the use/mention distinction, as well as treat vectors of variables as sets whenever this is required for meaningfulness.

An atomic sentence $X = x$ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, all solutions to the structural equations represented by $\mathcal{F}$ assign value x to the endogenous variable X if the exogenous variables in $\vec{\mathcal{U}}$ are set to $\vec{u}$. Since we are restricting the discussion to extended acyclic causal models which have a unique solution in any given context, this means that $X = x$ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, x is the value of X in the unique solution to all equations in $\mathcal{M}$ in $\vec{u}$. The truth conditions for negations, conjunctions, and disjunctions are given in the usual way.

[1] Gallow (2021: 59ff) ignores that removing an exogenous variable U from a causal model $\mathcal{M}$ does not, in general, result in a new causal model. The reason is that the endogenous variables whose only parent in $\mathcal{M}$ is U need not be the only variables that are left without any parent after U has been removed (e.g., if the actual value of U is 0 and the value of V_1 is determined by $V_2 \times U$).

The counterfactual $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \boxright \phi$, or simply $\vec{X} = \vec{x} \boxright \phi$, is true in $\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$ in $\vec{u}$ if, and only if, ϕ is true in $\mathcal{M}_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ in $\vec{u}$. The latter causal model results from $\mathcal{M}$ by removing the structural equation for X_i and by freezing the value of X_i at x_i, for each $i = 1, \ldots, k$. Formally, this means that $\mathcal{S}$ is reduced to $\mathcal{S}_{\vec{X}} = \langle \mathcal{U}, \mathcal{V} \backslash \{X_1, \ldots, X_k\}, \mathcal{R} \restriction_{\mathcal{U} \cup \mathcal{V} \backslash \{X_1, \ldots, X_k\}} \rangle$, where $\mathcal{R} \restriction_{\mathcal{U} \cup \mathcal{V} \backslash \{X_1, \ldots, X_k\}}$ is $\mathcal{R}$ with its domain restricted from $\mathcal{U} \cup \mathcal{V}$ to $\mathcal{U} \cup \mathcal{V} \setminus \{X_1, \ldots, X_k\}$; as well as that $\mathcal{F}$ is reduced to $\mathcal{F}^{\vec{X}=\vec{x}}$ which results from $\mathcal{F}$ by deleting, for each $i = 1, \ldots, k$, the function F_{X_i} representing the structural equation for X_i and by changing the remaining functions F_Y in $\mathcal{F} \setminus \{F_{X_1}, \ldots, F_{X_k}\}$ as follows: restrict the domain of each F_Y from $\times_{X \in \mathcal{U} \cup \mathcal{V} \backslash \{Y\}} R(X)$ to $\times_{X \in \mathcal{U} \cup \mathcal{V} \backslash \{Y, X_1, \ldots, X_k\}} R(X)$; and, replace F_Y by $F_Y^{\vec{X}=\vec{x}}$ which results from F_Y by setting $X_1, \ldots, X_k$ to $x_1, \ldots, x_k$, respectively.

The formal details of the structural equations semantics may look complicated, but the idea behind it is simple. In evaluating the counterfactual $\vec{X} = \vec{x} \boxright \phi$ in causal model $\mathcal{M}$ in context $\vec{u}$, first actively intervene and, all by yourself, cause the antecedent to be true by deleting the equations for the endogenous variables $\vec{X}$ and by subsequently setting the values of these variables to $\vec{x}$. In a second step, actively intervene and set the exogenous variables to $\vec{u}$; then let the remaining equations determine the values of the remaining endogenous variables. In a third step, check if the resulting solution yields the right value for ϕ.

The structural equations represent the "(causal) laws" or mechanisms of the model – what I have referred to as "causal structure" in section 8.3. These "laws" may fail to meet many of the traditional criteria for lawlikeness or "lawfulness" (Woodward 2003: ch. 6). In particular, structural equations differ from laws of nature in the sense of the best system analysis (section 8.3). Woodward (2003: ch. 7) characterizes structural equations as invariant under interventions, as well as modular, or autonomous: the structural equation for an endogenous variable is invariant under a range of interventions upon its parents; and, each endogenous variable can itself be intervened upon without affecting the structural equation for any other endogenous variable in the causal model. Essential to these ideas, which can be traced back to Haavelmo (1944), is the causal notion of an intervention.

Informally, an intervention – that is, a hard intervention – on a target variable in a causal model is a new variable in a new causal model that is directly causally relevant to the target variable, completely determines a specific value of the latter, and does so completely on its own, thereby erasing all previously existing relations of direct causal relevance into the target variable from the latter's previous parents. (For details see Woodward 2003: ch. 3; 2016, Woodward & Hitchcock 2003, and Hitchcock & Woodward 2003).

In contrast to this, a possibly soft intervention on a target variable in a causal model is a new variable in a new causal model that is directly causally relevant to the target variable, but merely constrains or influences which value the latter takes on – without necessarily completely determining a specific value of the target variable, as well as without necessarily erasing any previously existing relations of direct causal relevance into the target variable from the parents of the target variable (see Correa & Bareinboim 2020a, as well as Tian & Pearl 2001, Korb et al. 2004, Markowetz et al. 2005, Eaton & Murphy 2007, Eberhardt & Scheines 2007, Kocaoglu et al. 2019, Correa & Bareinboim 2020b, and Jaber et al. 2020).

In these informal terms, an interventionist counterfactual $\alpha \dashv\!\!\exists\ \gamma$ can be read as follows: if α were brought about by any collection of joint hard interventions on some or all of the variables in α, but no others, then γ would be the case. In contrast, our more general causal counterfactual $\alpha \boxdot\!\!\rightarrow \gamma$ is to be read as follows: if α, but no logically stronger proposition, were brought about by any collection of potentially soft interventions, then γ would be the case.

Interventionist counterfactuals with the structural equations semantics have antecedents of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$. For these, the two readings coincide and reduce to the following: the consequent would be the case, if the antecedent were brought about by the collection of k joint hard interventions on $X_1, \ldots, X_k$ that set their values to $x_1, \ldots, x_k$, respectively. The two readings come apart if we consider interventionist counterfactuals with antececents of a different form, as does Briggs (2012), who allows for arbitrary Boolean combinations of atomic sentences as antecedents (as well as arbitrary consequents).

To illustrate the difference between Galles & Pearl (1998)'s interventionist counterfactuals with a structural equations semantics in Halpern (2013)'s more general form, Briggs (2012)' interventionist counterfactuals with a possible states semantics, and our causal counterfactuals, let us consider Pearl (2000/2009: 27)'s very first example of a structural equations model: an econometric model relating unit price P and household demand Q for a given product to household income I and wage rate W for producing the product:

$$\begin{aligned} Q &= b_1 \cdot P + d_1 \cdot I + U_1 \\ P &= b_2 \cdot Q + d_2 \cdot W + U_2 \end{aligned}$$

The lower case letters are real numbers. U_1 and U_2 are so-called "error terms" that lump together all factors affecting household demand and unit price other than household income and wage rate. The variables are understood in a generic sense, as is common in the sciences. Specifically, the causal model and its associated directed graph (figure 9.3) are cyclic, as we are dealing with a feedback system.

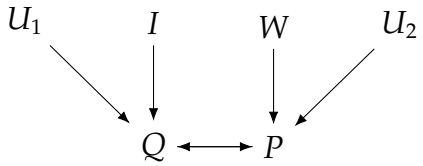

Figure 9.3: price and demand (generic variables)

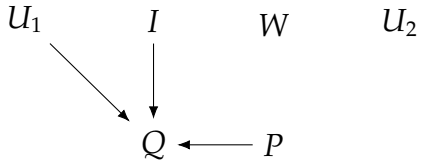

Figure 9.4: price and demand after a hard intervention on price

(As an aside, by proceeding as in section 9.1, except in reverse, we can render the causal model and its associated directed graph acyclic, as well as one with singular variables: we consider today's price P of a bottle of next year's DRC Romanée-Conti GC, tomorrow's demand Q by Ida for a bottle of this wine, Ida's present wealth I, and the cost W for producing a bottle of this wine. The structural equations are $Q = a \cdot P + b \cdot I + U_1$ and $P = c \cdot W + U_2$, and the associated directed acyclic graph is like figure 9.3, except that there is no arrow from Q to P.)

To illustrate the autonomy of these two structural equations, as well as the distinction between passively observing and actively intervening, manipulating, or controlling, Pearl (2000/2009: 28) asks us to consider that "government decides on price control and sets the price P at p_0." With the help of the two equations we can then compute the truth of the following generic interventionist counterfactual:

HIC If government were to decide on price control and set the price P at p_0, then demand Q would equal $b_1 \cdot p_0 + d_1 \cdot I + U_1$.

The hard intervention results in a new causal model and new associated directed graph (figure 9.4):

$$\begin{aligned} Q &= b_1 \cdot p_0 + d_1 \cdot I + U_1 \\ P &= p_0 \end{aligned}$$

Note that neither represents the intervention variable, as is common and in line with the structural equations semantics stated in the previous paragraphs. (This will be a feature of our definition of hard and soft interventions in section 9.8).

In some cases a government may be able to completely control the price of a product and set it to a specific value. However, even if we ignore black markets and people not following a government's policies in other ways, in many cases such hard interventions are merely hypothetical. For a number of reasons it is more common that governments decide to merely constrain or influence the value of a variable without setting it to a specific value. For instance, a government is much more likely to introduce a minimum wage or cap the allowable amount charged for a drug rather than set wages or prices to specific values. If they have any effect at all, these government actions correspond to soft interventions which give rise to counterfactuals such as the following:

SIC If government were to decide on price control and set the price P of 1 bottle of wine to *at least* 1 dollar (by means of a tax of 1 dollar per bottle of wine), then household demand Q would remain above 1 bottle of wine per week.

Since its antecedent is not a conjunction of one or more atomic sentences, SIC is not in the language for which the structural equations semantics is given. So, for this trivial reason, SIC is not assigned a truth value. On the possible states semantics, it is false. The reason is that, on this semantics, SIC is true if, and only if, every collection of joint hard interventions on some or all of the variables in the antecedent, but no others, that makes the antecedent true results in the truth of the consequent. This includes interventions that set the tax to arbitrarily high, but finite amounts of dollars – that is, interventions that amount to bans on wine. On our typicality semantics, SIC is true on some assignments of typicality and false on others. Whatever one thinks about SIC, the point is that it is not assigned a truth value by the structural equations semantics and means different things on Briggs (2012)' possible states semantics, which follows Galles & Pearl (1998) and Halpern (2000; 2013) in restricting itself to hard interventions, and our typicality semantics which allows for hard and soft interventions.

A hard intervention deletes the structural equation for an endogenous variable, freezes the latter at a specific value, and removes all previously existing arrows into it. The soft intervention just considered still works with just the existing structural equation. Instead of deleting the structural equation for price, though, it merely constrains the value of the latter. Therefore, the arrows into price cannot be removed anymore: wage rate and the other factors still affect price. There are soft interventions that do not work with just the existing structural equation for the target variable. Some change the structural equation. Furthermore, they do so while removing all, some, or none of the previously existing arrows, as well as while possibly introducing new arrows.

Soft interventions that do not work with just the existing structural equation can be illustrated by government policies that change a mechanism. For instance, having children costs money. A government may decide to cover all of these costs. In this case a previously existing causal relationship between having children and one's finances disappears: an arrow is removed. A government may also decide to cover all of these costs plus pay a bonus. In this case having children still affects one's finances, but in the opposite way: an arrow is not removed, and while the arrow itself still means the same thing – that is, direct causal relevance – the new structural equation that gives rise to it represents a very different mechanism.

Alternatively, a government may decide to criminalize certain behavior so that now there are consequences that previously did not exist: a new arrow between previously existing nodes representing previously existing variables is introduced. All of these soft interventions give rise to counterfactuals that are meaningless on both the structural equations and Briggs (2012)' possible states semantics. The reason is that their antecedents involve not only Boolean combinations of atomic sentences, but also counterfactuals. These counterfactuals are meaningful on our typicality semantics if we allow for iterations of typicality (see section 7.4). In fact, while we will not do so, by proceeding as in section 6.1, one may even be able to capture interventions that introduce new variables into a causal model.

There is some debate about whether Pearl (2000/2009)'s *do*-operator, which appeals to hard interventions only, is able to capture all aspects of intervening. At least when it comes to causal inference, Cartwright (2007) and Heckman & Pinto (2015) (see also Heckman 2005) are less optimistic than Pearl (2010) and Pearl (ms), respectively (see also Pearl 2021: 432). We are presently focusing on the semantics of causal counterfactual and will turn to inference only in the next chapter. In this case, one may hold the view that the concept of an intervention simply does not make sense in its soft form: to be meaningful, an intervention must set the variable intervened upon to a specific value – even if, in practice and, hence, in relation to inference, one may sometimes be able to carry out only a soft intervention that merely constrains the value of the variable intervened upon.

On this view of interventions, the results of this chapter are of limited interest, though not none: the interventionist counterfactuals with the structural equations semantics are reproduced in the typicality semantics; in addition, their difference to backtracking counterfactuals is characterized. Briggs (2012)' possible states semantics seems to do more, as it does not merely reproduce these interventionist counterfactuals, but defines the possible states semantics for a richer language. However, this generalization does not really go beyond Galles & Pearl (1998)'s original in the form it takes in Halpern (2013), as I will try to explain now.

Briggs (2012) allows for interventionist counterfactuals with arbitrary Boolean combinations of atomic sentences as antecedents, as well as arbitrary consequents, including interventionist counterfactuals. In addition, Briggs (2012) allows for Boolean embeddings. Alas, it follows from a result by Briggs (2012: 163) that each sentence in this language is exactly equivalent to one in Halpern (2013)'s. Exact equivalence is quite strict: for Boolean combinations of atomic sentences to which the truth conditions of classical logic apply, strict equivalence implies, but is not implied by logical equivalence. Like Briggs (2012), Halpern (2013) allows for Boolean embeddings, but continues to require that interventionist counterfactuals have non-empty conjunctions of atomic sentences as antecedents and Boolean combinations of atomic sentences as consequents. The reason for the unfortunate consequence is that, on Briggs (2012)' view, interventionist counterfactuals with disjunctive or negated antecedents have exactly the same meaning as *con*junctions of interventionist counterfactuals whose antecedents are conjunctions of atomic sentences and whose consequents are Boolean. Interventions on disjunctions and negations are construed as *con*junctions of collections of joint hard interventions. If one insists on generalization, then one needs to generalize the very notion of an intervention itself. Otherwise, the logic of interventionist counterfactuals is completely characterized by Galles & Pearl (1998), modulo Halpern (2000; 2013).

Even Briggs (2012) restricts the antecedents of interventionist counterfactuals to Boolean combinations of atomic sentences, and I would like to conclude this section with an informal diagnosis of why. Consider a causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$ and a model $\langle W, (\$_w)_{w \in W} \rangle$ in the traditional possible worlds semantics, where, for any possible world w in W, $\$_w$ is, say, an accessibility relation over W that we interpret as representing w's laws. The set of possible worlds $\mathcal{W} = \times_{X \in \mathcal{U} \cup \mathcal{V}} R(X)$ of the causal model contains some that satisfy the structural equations represented by $\mathcal{F}$; the others, called "illegal" in Glymour et al. (2010), do not. The illegal possible worlds are not merely illegal, though, but outright *lawless*: they violate at least one of the structural equations represented by $\mathcal{F}$, but the causal model does not specify whether they satisfy alternative ones. It is for this reason that interventionist counterfactuals do not have truth values at illegal possible worlds. By contrast, the classical truth conditions for Boolean combinations of atomic sentences still apply at them. This in turn is why we can consider as antecedents of interventionist counterfactuals the latter sentences, but not the former. A model in the traditional possible worlds semantics specifies, for *each* possible world w, w's laws. In contrast to this, a causal model specifies the causal laws of merely the actual world (and all legal ones). This incompleteness in its modal component $\mathcal{F}$ prevents interventionist counterfactuals from having arbitrary antecedents.

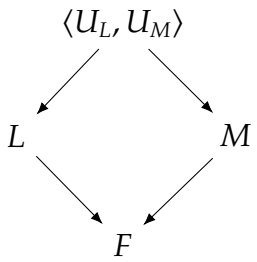

Figure 9.5: fire

9.4 Normality

Halpern & Hitchcock (2010)'s definition renders actual causation relative to a causal model, as well as the default values for the endogenous variables of the latter. To illustrate, let us consider their fire example. Endogenous variable L takes on value 1 if there is lightning, and 0 otherwise. Endogenous variable M takes on value 1 if there is an arsonist dropping a lit match, and 0 otherwise. Endogenous variable F takes on value 1 if there is a forest fire, and 0 otherwise. Furthermore, exogenous variable $\langle U_L, U_M\rangle$ directly causally determines the value of L and M. The functions $F_L : \langle\langle i, j\rangle, m, f\rangle \mapsto i$, $F_M : \langle\langle i, j\rangle, l, f\rangle \mapsto j$, and $F_F : \langle\langle i, j)\rangle, l, m\rangle \mapsto \max\{l, m\}$ describe the following structural equations which are pictured by a directed acyclic graph (figure 9.5):

$$L = U_L$$

$$M = U_M$$

$$F = L \vee M$$

I will follow the common practice of identifying the functions $1-$, $\min$, and $\max$, respectively, by the logical connectives $\neg$, $\wedge$, and $\vee$, respectively, if all variables are binary. I will follow also the common practice of writing down structural equations as equations, even though this is not what they are.

As Halpern & Hitchcock (2010) explain, in the context where $U_L = 1$ and $U_M = 1$ so that there is lightning ($L = 1$), an arsonist dropping a lit match ($M = 1$), and a forest fire ($F = 1$), the arsonist's dropping a lit match ($M = 1$) is an actual cause of the forest fire ($F = 1$). The reason is the following, where, to keep things readable, the rank of a sentence is identified with the rank of the set of possible worlds in which the sentence is true and the rank of a possible world is identified with the rank of the singleton containing it.

1. $M = 1$ and $F = 1$ are true in $\mathcal{M}$ in $\langle 1, 1\rangle$;

2. for the partition $\{\{M, F\}, \{L\}\}$ and the values 0 and 0 of M and L we have $\varrho_{\langle 1,1\rangle}(M = 0 \wedge L = 0) \leq \varrho_{\langle 1,1\rangle}(w_{\langle 1,1\rangle})$, as well as: $\langle M, F\rangle = \langle 1, 1\rangle$ is true in $\mathcal{M}$ in $\langle 1, 1\rangle$, and so are

   (a) $M = 0 \wedge L = 0 \mathrel{-\!\!\!\wedge} F \neq 1$,

   (b) $M = 1 \wedge L = 0 \wedge F = 1 \mathrel{-\!\!\!\wedge} F = 1$, $M = 1 \wedge L = 0 \mathrel{-\!\!\!\wedge} F = 1$, $M = 1 \wedge F = 1 \mathrel{-\!\!\!\wedge} F = 1$, and $M = 1 \mathrel{-\!\!\!\wedge} F = 1$; and

3. there is no proper subset of $\{M\}$ such that 1. and 2. hold.

The inequality for the ranking function $\varrho_{\langle 1,1\rangle}$ says that the least atypical possible worlds where there is no lightning and no arsonist dropping a lit match are at most as atypical as the actual world where there is lightning, an arsonist dropping a lit match, and a forest fire. This inequality holds (in the context where $U_L = 1$ and $U_M = 1$) because it is more typical that there is no lightning than that there is lightning, that there is no arsonist dropping a lit match than that there is an arsonist dropping a lit match, and that there is no forest fire than that there is a forest fire.

In addition, the structural equations seem to put a constraint on the assignment of normality or typicality. Even though it is more typical that there is no forest fire than that there is a forest fire, it is more typical that there is lightning and a forest fire than that there is lightning and no forest fire. Similarly, even though it is more typical that there is no forest fire than that there is a forest fire, it is more typical that there is an arsonist dropping a lit match and a forest fire than that there is an arsonist dropping a lit match and no forest fire. Finally, even though it is more typical that there is no forest fire than that there is a forest fire, it is much more typical that there is lightning and an arsonist dropping a lit match and a forest fire than that there is lightning and an arsonist dropping a lit match, but no forest fire. And this is so no matter which context we are in.

More generally, the structural equations seem to put the following constraint on the assignment of normality or typicality. Illegal possible worlds which violate at least one structural equation are less typical than legal possible worlds that obey all structural equations. Furthermore, illegal possible worlds which violate certain structural equations and then some are less typical than illegal possible worlds which violate only certain structural equations. That is, a possible world that violates the structural equations for a set of endogenous variables is less typical than a possible world that violates the structural equations for only a proper subset of this set.

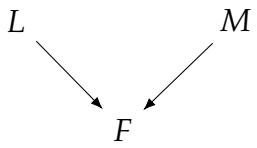

Figure 9.6: fire (simplified version)

Our constraint is meaningful for a quantitative concept of normality such as our rank-theoretic notion of typicality (section 7.4), as well as a comparative one that, besides being reflexive and transitive, may be connected (see section 8.3), but need not be (Halpern & Hitchcock 2013; 2015 and Halpern 2016). Halpern & Hitchcock (2013: 1004f) formulate the first part of this constraint in their Default Rule 1, which is supposed to hold by default only, though, not universally. Halpern & Hitchcock (2015: 435) and Halpern (2016: 81) implicitly subscribe to a related constraint (see section 9.6).

Our constraint does not hold for the structural equations $L = U_L$ and $M = U_M$, if only because we have not even specified what the exogenous variables U_L and U_M mean. It would be a mistake to hold this against our constraint, though. Rather, it should be taken as reason to reject the causal model of the fire example. Halpern & Hitchcock (2010) include U_L, U_M, $L = U_L$, and $M = U_M$ only because they want to say that $L = 1$ and $M = 1$ are actual causes of $F = 1$, but cannot do so unless L and M are endogenous variables. Besides that, these exogenous variables and structural equations do no work: they could be dropped without loss if the restriction were not in place that only endogenous variables can be causally efficacious. If that restriction were not in place, L and M would be exogenous and $F = L \vee M$ the only structural equation. Indeed, this would be the causal model in Hitchcock (2007)'s framework. Furthermore, in Halpern & Hitchcock (2015) and Halpern (2016) these redundant exogenous variables and structural equations are generally omitted from the directed acyclic graphs picturing (extended acyclic) causal models.

Without this restriction, the fire example is both simpler and more natural. Let *exogenous* variable L take on value 1 if there is lightning, and 0 otherwise. Let *exogenous* variable M take on value 1 if there is an arsonist dropping a lit match, and 0 otherwise. Let endogenous variable F take on value 1 if there is a forest fire, and 0 otherwise. The function $F_F : \langle l, m\rangle \mapsto \max\{l, m\}$ describes the following structural equation which is pictured by a directed acyclic graph (figure 9.6):

$$F = L \vee M$$

In this causal model possible worlds which violate a structural equation are less typical than possible worlds which obey all structural equations. Therefore, my first proposal is to drop the aforementioned restriction in the (extended acyclic) causal models of Halpern & Pearl (2005a), Halpern (2008; 2016), and Halpern & Hitchcock (2010; 2013; 2015) and define an atomic sentence to be of the form $X = x$ for an exogenous or endogenous variable X in $\mathcal{U} \cup \mathcal{V}$ and a value x in its range $R(X)$. Then we do not have to include arbitrary exogenous variables to render L and M endogenous and, thus, be able to state counterfactual and causal claims with them.

For this to make sense we have to define the truth conditions for sentences in a slightly different way. An atomic sentence $X = x$ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, all solutions to the structural equations represented by $\mathcal{F}$ assign value x to the exogenous or endogenous variable X if the exogenous variables are set to $\vec{u}$. Since we keep restricting the discussion to (extended) acyclic causal models which have a unique solution in any context, this means that $X = x$ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, x is the value of X in the unique solution to all equations in $\mathcal{M}$ in $\vec{u}$. The truth conditions for negations, conjunctions, and disjunctions are again given in the usual way. The interventionist counterfactual $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \boxminus \phi$ is true in $\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$ in $\vec{u}$ if, and only if, ϕ is true in $\mathcal{M}_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ in $\vec{u}_{\vec{X}=\vec{x}}$.

The new causal model $\mathcal{M}_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ results from old the causal model $\mathcal{M}$ by removing the structural equations for the endogenous variables among $X_1, \ldots, X_k$ and by freezing the value of X_i at x_i, for each $i = 1, \ldots, k$. Formally, this means that $\mathcal{S}$ is reduced to $\mathcal{S}_{\vec{X}} = \langle \mathcal{U}, \mathcal{V} \setminus \{X_1, \ldots, X_k\}, \mathcal{R} \restriction_{\mathcal{U} \cup (\mathcal{V} \setminus \{X_1, \ldots, X_k\})} \rangle$, where $\mathcal{R} \restriction_{\mathcal{U} \cup (\mathcal{V} \setminus \{X_1, \ldots, X_k\})}$ is $\mathcal{R}$ with its domain restricted from $\mathcal{U} \cup \mathcal{V}$ to the set of variables $\mathcal{U} \cup (\mathcal{V} \setminus \{X_1, \ldots, X_k\})$ which remain after deleting the endogenous variables among $X_1, \ldots, X_k$; and that $\mathcal{F}$ is reduced to $\mathcal{F}^{\vec{X}=\vec{x}}$ which results from $\mathcal{F}$ by deleting the functions among $F_{X_1}, \ldots, F_{X_k}$ representing the structural equations for the endogenous variables among $X_1, \ldots, X_k$ and by changing the remaining functions F_Y in $\mathcal{F} \setminus \{F_{X_1}, \ldots, F_{X_k}\}$ as follows: restrict the domain of each F_Y from $\times_{X \in \mathcal{U} \cup \mathcal{V} \setminus \{Y\}} R(X)$ to $\times_{X \in \mathcal{U} \cup (\mathcal{V} \setminus \{Y, X_1, \ldots, X_k\})} R(X)$ and replace F_Y by $F_Y^{\vec{X}=\vec{x}}$ which results from F_Y by setting $X_1, \ldots, X_k$ to $x_1, \ldots, x_k$, respectively.

The new context $\vec{u}_{\vec{X}=\vec{x}}$ results from the old context $\vec{u}$ by setting the values of the exogenous variables among $X_1, \ldots, X_k$ to $x_1, \ldots, x_k$, respectively, and by leaving the values of the other exogenous variables in $\mathcal{U} \setminus \{X_1, \ldots, X_k\}$ as they are in $\vec{u}$.

The definition of actual causation has to be changed slightly. In clause 2. we consider a partition of the set $\mathcal{U} \cup \mathcal{V}$ of all variables, exogenous and endogenous, rather than a partition of the set $\mathcal{V}$ of endogenous variables only.

Let us apply this definition to Halpern & Hitchcock (2010: 400)'s survival example which, in tandem with their fire example in its simplified form, shows the need for normality (Halpern 2016: 88ff argues that normality may not be needed to deal with this example after all, but that it helps with a variant thereof, though not with other examples). Let exogenous variable A take on value 1 if Assassin does not put in poison, and 0 otherwise. Let exogenous variable B take on value 1 if Bodyguard puts in antidote, and 0 otherwise. Let endogenous variable S take on value 1 if Victim survives, and 0 otherwise. The function $F_S : \langle a, b\rangle \mapsto \max\{a, b\}$ describes the following structural equation which is pictured by a directed acyclic graph (figure 9.7):

$$S = A \vee B$$

Except for the meaning of the variables, the structural equation for the survival example is identical to the structural equation for the fire example (in its simplified form). In addition, in both cases all variables take on value 1. Yet, according to Hitchcock (2007), Halpern (2008; 2016), and Halpern & Hitchcock (2010; 2015), the correct causal judgments for these two examples differ. In the fire example, the arsonist's dropping a lit match is an actual cause of the forest fire if there is lightning and an arsonist dropping a lit match and a forest fire. By contrast, in the survival example, it is not the case that Bodyguard's putting in antidote ($B = 1$) is an actual cause of Victim's survival ($S = 1$) if Bodyguard puts in antidote ($B = 1$) and Assassin does not put in poison ($A = 1$) and Victim survives ($S = 1$). This difference in causal judgments is explained by appeal to normality. While the structural equation for the two examples is identical except for the meaning of the variables, and so are the values the variables take on, the ordering of normality in the two examples differs.

In the context where Assassin does not put in poison and Bodyguard puts in antidote, it is more typical that Assassin does not put in poison than that Assassin puts in poison, that Bodyguard does not put in antidote than that Bodyguard puts in antidote, and that Victim survives than that Victim does not survive. In addition, the structural equation seems to put a constraint on the assignment of normality or typicality. Even though it is more typical that Victim survives than that Victim does not survive, it is more typical that Assassin puts in poison and Bodyguard does not put in antidote and Victim does not survive than that Assassin puts in poison and Bodyguard does not put in antidote and Victim survives.

This helps us see why it is not the case that Bodyguard's putting in antidote is an actual cause of Victim's survival if Bodyguard puts in antidote, Assassin does not put in poison, and Victim survives.

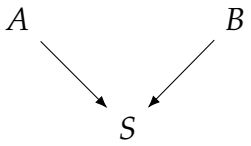

Figure 9.7: survival

1. $B = 1$ and $S = 1$ are true in $\mathcal{M}$ in $\langle 1, 1\rangle$; but

2. for the partition $\{\{B, S\}, \{A\}\}$ (and any other partition) there are no values b and a of B and A with $\varrho_{\langle 1,1\rangle}(B = b \wedge A = a) \leq \varrho_{\langle 1,1\rangle}(w_{\langle 1,1\rangle})$, as well as: $\langle B, S\rangle = \langle 1, 1\rangle$ is true in $\mathcal{M}$ in $\langle 1, 1\rangle$, and so are

   (a) $B = b \wedge A = a \mathrel{-\!\!\!\exists} S \neq 1$,

   (b) $B = 0 \wedge A = a \wedge S = 1 \mathrel{-\!\!\!\exists} S = 1$, $B = 1 \wedge A = a \mathrel{-\!\!\!\exists} S = 1$, $B = 1 \wedge S = 1 \mathrel{-\!\!\!\exists} S = 1$, and $B = 1 \mathrel{-\!\!\!\exists} S = 1$; and

3. there is no proper subset of $\{B\}$ such that 1. and 2. hold.

The reason is that the values b and a of B and A needed for $B = b \wedge A = a \mathrel{-\!\!\!\exists} S \neq 1$ in part (a) of clause 2. to come out true in $\mathcal{M}$ in $\langle 1, 1\rangle$ are 0 and 0. However, every possible world in which Bodyguard does not put in antidote and Assassin puts in poison, that is, where $B = 0 \wedge A = 0$ is true, is less typical than the actual world $w_{\langle 1,1\rangle}$ where Bodyguard puts in antidote and Assassin does not put in poison – or so Halpern & Hitchcock (2010: sct. 5) claim.

In fact, however, this is not true for the ranking function used by Halpern & Hitchcock (2010) which assigns rank 1 to both the possible world that would be needed where Bodyguard does not put in antidote and Assassin puts in poison and the actual world where Bodyguard puts in antidote and Assassin does not put in poison. What is true, though, is that the possible world that would be needed where Bodyguard does not put in antidote and Assassin puts in poison is less typical than *the maximally typical possible world* where Assassin does not put in poison, that is, the possible world where Bodyguard does not put in antidote and Assassin does not put in poison.

Therefore, one way to fix this minor bug is to adjust the definition of actual causation (in the spirit of Hitchcock 2007, who also refers to the actual value of $\vec{W}$ rather than the actual world) as follows: in clause 2., $\varrho_{\vec{u}}\left(\vec{X} = \vec{x'} \wedge \vec{W} = \vec{w}\right) \leq \varrho_{\vec{u}}\left(\vec{W} = \vec{w_{\vec{u}}}\right)$, where $\vec{w}_{\vec{u}}$ is the actual value of $\vec{W}$ in causal model $\mathcal{M}$ in context $\vec{u}$.

An alternative way, preferred by Halpern & Hitchcock (2015) and Halpern (2016: ch. 3), is to stick to the original definition, but stipulate that the unique possible world that is determined by the context and the interventions that set the values of A and B to 0 and 0 is not at least as typical as the actual world. Rather, these two possible worlds cannot be compared with respect to their normality (in this context).

9.5 Typicality

This finally brings us to the formal representation of normality, as well as its philosophical interpretation. Like Hitchcock & Knobe (2009) and Bear & Knobe (2017), Halpern & Hitchcock (2010; 2013; 2015) and Halpern (2016) interpret normality as including both purely descriptive and evaluative elements. Unlike Hitchcock & Knobe (2009) (see section 8.2), Sytsma et al. (2012), Bear & Knobe (2017), and Morris et al. (ms), Halpern (2008; 2016) and Halpern & Hitchcock (2010; 2013; 2015) understand normality in a singular, not generic sense – and must do so to be able to construct the set of possible worlds in the way outlined in section 9.1. So, we are dealing with a singular concept of all-things-considered normality that includes both purely descriptive and evaluative elements – unlike our singular, but purely descriptive concept of typicality from chapter 7.

One possibility is that all-things-considered normality results by aggregating various individual aspects of normality some of which are purely descriptive and some of which are evaluative. In this case we face the same obstacles as with the aggregation of individual aspects of similarity to overall similarity (section 8.3), if normality is represented by an ordering relation that is reflexive, transitive, and connected. Unfortunately, as Weymark (1984) and Sen (1986) show, these obstacles persist even if normality is represented by a mere quasi-ordering that is reflexive and quasi-transitive (that is, its strict part "strictly more normal than" is transitive), but not necessarily connected or antisymmetric (or even transitive), such as the pre-ordering of Halpern & Hitchcock (2013; 2015) and Halpern (2016: ch. 3). The good news is that, as mentioned in section 8.3, these obstacles can be overcome by representing the aspects that are to be aggregated numerically by a ranking function. In particular, if the purely descriptive aspects of normality are represented by some ranking functions and the evaluative ones by others, then one may combine them to all-things-considered normality that is represented by a ranking function r or, if one prefers, a comparative pre-ordering that is reflexive and transitive, but not necessarily connected or antisymmetric.[2]

Consequently, on this possibility, there are reasons to represent the individual aspects of normality by ranking functions. Prima facie, there are reasons also to think that this possibility obtains. Together with a specification of the values of all (exogenous) variables, the structural equations of an extended (acyclic) causal model determine what occurs, what depends counterfactually on what, and which interventions would bring about which outcomes. According to Hitchcock (2007), Halpern (2008; 2016), and Halpern & Hitchcock (2010; 2013; 2015), normality plays a role in selecting, from among all conditions on which an effect depends counterfactually (in some appropriate sense; see section 9.2), those that are its actual causes. According to Hitchcock & Knobe (2009) and Morris et al. (ms), actual causation plays a role in selecting, from among all interventions that would bring about a desired outcome, those that are most effective or, perhaps, preferable for other reasons. Prima facie, which intervention is preferable may depend on both purely descriptive and evaluative considerations. So, prima facie, normality includes both purely descriptive and evaluative elements. Hitchcock & Knobe (2009) identify statistical norms, moral norms, and norms of proper functioning among the senses of normality determining which intervention is preferable. In section 8.2 I argued that these need to be understood in a singular, not generic sense, if we talk about actual causation between events or related relata. Now I will argue that they need to be understood in a purely descriptive, not evaluative sense.

First, though, note that there are concepts involving both purely descriptive and evaluative elements that play useful roles in comparable situations. Consider decisions under objective risk (as opposed to subjective uncertainty). When faced with such a decision, we have to select, from among all available acts, those that are preferable. According to classical (normative) decision theory, we (ought to) do so by considering the purely descriptive chance of each possible state of the world and the evaluative subjective utility / objective value of the outcome of any available act in any possible state of the world. Then we (ought to) aggregate the purely descriptive and evaluative components by forming the expected subjective utilities / objective values of the available acts – another singular, not generic concept, as the acts are to be understood as tokens, not types. Finally, we (ought to) select the acts that maximize this aggregate.

[2] A relation $\preceq$ among possible worlds is antisymmetric if, and only if, for all possible worlds v and w: if $v \preceq w$ and $w \preceq v$, then $v = w$. A pre-ordering of normality among possible worlds that is reflexive and transitive, but not necessarily connected or antisymmetric, can be obtained from a rank-theoretic normality function r as follows: v is as normal as w if, and only if, $r(v) = r(w)$; v is more normal than w if, and only if, $r(v) < r(w) + k$, for a fixed natural number k.

Prima facie, normality plays a role in the selection of interventions that is somewhat similar to the role played by expected subjective utility / objective value in the selection of acts. Just as which acts are preferable depends on the purely descriptive chances of the possible states of the world, so too which interventions are preferable depends on how likely they are to succeed. Just as which acts are preferable depends on the evaluative subjective utilities / objective values of the outcomes of the acts in the states of the world, so too which interventions are preferable depends on how subjectively costly / objectively wrong it is to carry them out.

To illustrate, the forest fire in the fire example is an undesirable outcome. An arsonist's dropping a lit match is among its actual causes. According to the reasoning outlined in the previous paragraphs, we should intervene on the variable M and set it from its actual value 1 to its default value 0. This makes perfect sense. Attempting to prevent the forest fire by intervening and making it the case that no arsonist drops a lit match is an excellent strategy to prevent the forest fire. The same is true for lightning, which too is an actual cause of the forest fire.

The sense in which lightning is abnormal is purely descriptive. According to Hitchcock & Knobe (2009: 605ff), the reason it makes sense to intervene on the variable L and set it from its actual value 1 to its default value 0 is that, so to speak, typically, things are typical. Therefore, the intervention is, in a purely descriptive and singular sense, likely to succeed. There is no logical guarantee that this intervention prevents the forest fire, as atypical things may be the case. For instance, there may be someone who would set the forest on fire if, and only if, lightning were to strike. However, typically, this intervention does succeed.

One may object that this argument ignores the structural equation $L = F \vee M$ which stipulates, among others, that there would be a forest fire if there were lightning. It is impossible, so the objection, for the intervention to be carried out, but the forest fire not be prevented – and this is so no matter whether we interpret the variables in a singular or generic sense. One way one may respond is to point out that it is not impossible for the intervention to succeed in preventing the forest fire. It is merely illegal and, hence (if one subscribes to our constraint), atypical. In addition, perhaps Hitchcock & Knobe (2009)'s idea is not that the intervention is more likely to succeed in preventing the forest fire, but that it is more likely that one successfully carries out the intervention (which then prevents the forest fire with "legal necessity"). Arguably, it is easier to bring about a default state than some deviant state. After all, the former obtains by default. So, we can grant that an intervention that is selected on the basis of these purely descriptive considerations is more effective in at least some sense.

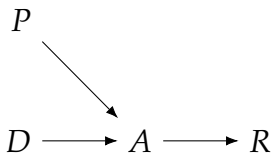

Figure 9.8: drug

The sense in which an arsonist's dropping a lit match is abnormal is certainly also descriptive, but perhaps not purely so: perhaps it is subjectively undesirable / objectively wrong for an arsonist to drop a lit match for reasons other than that it is an actual cause of the forest fire. Suppose so. According to Hitchcock & Knobe (2009: 605ff), it makes sense to intervene on the variable M and set it from its actual value 1 to its default value 0 independently of how, in a purely descriptive sense, likely this intervention is to succeed (or one is to succeed in carrying out this intervention). The reason is that this is so by stipulation: apart from likely preventing the forest fire, the successfully carried out intervention itself is subjectively more desirable / objectively better. Arguably, and as assumed implicitly by Hitchcock & Knobe (2009: 605ff), if the intervention is better in the purely descriptive sense, as well as the evaluative one, then it is also better overall.

So far, so good. Now consider an actual cause of a desirable effect, such as in Hitchcock & Knobe (2009: 602ff)'s drug example. Let exogenous variable P take on value 1 if Pharmacist signs off on Assistant's request for the drug, and 0 otherwise. Let exogenous variable D take on value 1 if Doctor signs off on Assistant's request for the drug, and 0 otherwise. Let endogenous variable A take on value 1 if Assistant administers the drug, and 0 otherwise. Let endogenous variable R take on value 1 if Patient recovers, and 0 otherwise. The functions $F_A : \langle p, d\rangle \mapsto \min\{p, d\}$ and $F_R : a \mapsto a$ describe the following structural equations which are pictured by a directed acyclic graph (figure 9.8):

$$A = P \wedge D$$

$$R = A$$

Pharmacist signs off on Assistant's request for the drug, as is typical. Doctor does so, too, as is abnormal because the drug is dangerous for patients like Patient. As it happens, Patient recovers and the Doctor's signing off is an actual cause thereof. According to Hitchcock & Knobe (2009: 608), "once again it makes sense to target the abnormal condition for intervention."

$$T \longrightarrow S$$

Figure 9.9: boat

Let us have a closer look. First, this intervention is unlike any we have come across so far. We are to intervene on the variable D, but instead of setting it to an alternative value, we leave it at its actual value. Instead, we are to change the assignment of normality for its values: the intervention is not to make Doctor do what Doctor did not do; the intervention is to make it normal for Doctor to do what previously was abnormal for Doctor to do. Call this a "second-order intervention."

We can grant that, as in the fire example, this makes sense for the purely descriptive and singular sense of normality. For instance, the hospital can adopt a new policy that recommends the drug for patients like Patient, thus making it, in a purely descriptive and singular sense, typical for Doctor to sign off. However, what does not make sense anymore is to second-order intervene on a variable and change which of its values are normal in an evaluative sense.

To see why, consider a desirable effect of an actual cause that is subjectively undesirable / objectively wrong, as in the following version of the trolley problem (Foot 1967). Exogenous variable T takes on value 1 if a person is thrown out of the boat, and 0 otherwise. Endogenous variable S takes on value 1 if the boat sinks, and 0 otherwise. The function $F_S : t \mapsto 1 - t$ describes the following structural equation which is pictured by a directed acyclic graph (figure 9.9):

$$S = \neg T$$

As desired, the boat does not sink. Its actual cause is the subjectively undesirable / objectively wrong fact that a person is thrown out of the boat.

To bring about the desirable effect, Hitchcock & Knobe (2009) have us second-order intervene on the variable T and make it subjectively desirable / objectively right to throw out a person for reasons other than that the boat does not sink. This makes no sense. While one may be able to change which values a variable, in a purely descriptive and singular sense, typically takes on, just as one may be able to change the chances with which a variable takes on its values, one cannot change which values of a variable are intrinsically valuable. This is so whether we have in mind subjective intrinsic desire or objective intrinsic value. What one may be able to do is change which values of a variable are extrinsically valuable. However, this is done by, not a second-order intervention on a given variable, but a soft first-order intervention on other variables that are – or, as a result of this intervention, come to be – causally downstream of the given variable.

This leaves norms of proper functioning. I assume these to be determined by background causal structure and, perhaps, statistical norms. Once again, consider an example with a desirable rather than undesirable effect. So, once again, we are to second-order intervene and make normal what previously was abnormal rather than first-order intervene and set the value of a variable to an alternative value. In this case we have to work with soft second-order interventions. The latter change, say (Hitchcock & Knobe 2009: 609ff), the workings of a machine or the set-up of a company's production line or whatever background causal structure is implicitly assumed to determine proper functioning, as opposed to the foreground causal structure explicitly stated by the structural equations of an extended causal model. We will see in the next section that structural equations and, hence, the causal structure they determine reduce to typicality in the purely descriptive sense (plus the causal, but purely descriptive distinction between exogenous and endogenous variables). So, this case has already been dealt with, except that soft second-order interventions, if they are spelled out explicitly and specify an alternative background causal structure, require typicality to be iterated.

To wrap up, I embrace the idea that normality, through its role in the selection of actual causes, plays a useful role in the selection of interventions that are, in a purely descriptive and singular sense, most efficient in at least one of two senses. I reject the idea that normality, through its role in the selection of actual causes, plays a useful role in the selection of interventions that are preferable for reasons which involve intrinsic value. Through its role in the selection of actual causes, normality plays a role in the selection of interventions that is similar to the role in the selection of acts played by, not expected utility / value, but chance. You inform me of what would be the actual causes of a potential effect. This is useful information on the basis of which I decide to intervene or second-order intervene, depending on my desire to prevent or bring about the potential effect. Like chance and typicality, actual causation is a purely descriptive and singular concept. To the extent that the evaluative elements of normality serve a purpose in the selection of interventions, they do so as mere means to indicate the presence of purely descriptive typicality: in this role, Hitchcock & Knobe (2009)'s moral norms, like their norms of proper functioning, are but a proxy for their statistical norms.

This flies in the face of the empirical results in Hitchcock & Knobe (2009) and Bear & Knobe (2017), and I owe an account of this tension. What is going on, I suggest, is the same that is going on in the Wason selection task (Wason 1966), the base rate fallacy (Bar-Hillel 1980 and Tversky & Kahneman 1982a), and the conjunction fallacy (Tversky & Kahneman 1982b; 1983): people systematically make mistakes (which does not automatically mean that they behave irrationally).

The documented performances in the Wason selection task are attributed to systematic mistakes in logical reasoning, not mistakes in classical logic. The base rate and conjunction fallacies are attributed to systematic mistakes in probabilistic reasoning, not mistakes in the probability calculus. I suggest we likewise attribute the empirical findings in Hitchcock & Knobe (2009) and Bear & Knobe (2017) to systematic mistakes in philosophical reasoning or judgment, not mistakes in the philosophical distinction between fact and value: this distinction is ignored by judgments of actual causation and normality that are sensitive to considerations of what is intrinsically valuable. I suggest we go even further and adopt this position not only with respect to judgments of actual causation and normality, but causal judgments more broadly. As a search on the website of the NY Times reveals, people frequently write that they are doing something "for a good cause." However, whether good, bad, or neither, the "causes" for which people are doing something are not causes, but ends or effects. Once again, people systematically make mistakes in philosophical reasoning or judgment, mixing up not only what is and what ought to be, but also what is cause and what effect.

You may not be particularly fond of my suggestion. Therefore, I hasten to add: nothing in what follows depends on it!

Consider an extended acyclic causal model $\langle \mathcal{S}, \mathcal{F}, (n_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ with a family $(n_{\vec{u}})_{\vec{u} \in R(\mathcal{U})}$ of formal representations n of all-things-considered normality, one for each context $\vec{u}$. Now consider $\langle \mathcal{S}, \mathcal{F}, (t_{\vec{u}})_{\vec{u} \in R(\mathcal{U})}, (n_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. $(t_{\vec{u}})_{\vec{u} \in R(\mathcal{U})}$ is a family of rank-theoretic typicality functions t such that, for each context $\vec{u}$, $t_{\vec{u}}$ represents the purely descriptive elements of $n_{\vec{u}}$. If, as in the possibility considered in the previous paragraphs, $n_{\vec{u}}$ results by aggregating the purely descriptive elements $t_{\vec{u}}$ and the evaluative ones $e_{\vec{u}}$, then one may consider $\langle \mathcal{S}, \mathcal{F}, (t_{\vec{u}})_{\vec{u} \in R(\mathcal{U})}, (e_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. If not – say, because the whole $n_{\vec{u}}$ is greater than the sum of its parts $t_{\vec{u}}$ and $e_{\vec{u}}$ – one has to consider $\langle \mathcal{S}, \mathcal{F}, (t_{\vec{u}})_{\vec{u} \in R(\mathcal{U})}, (n_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. Either way, now drop the right-most element to arrive at the extended acyclic causal model, restricted to its purely descriptive elements: $\langle \mathcal{S}, \mathcal{F}, (t_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. My claims are restricted to this restriction. Specifically, the constraint the structural equations "seem" to impose on the assignment of normality constrains purely descriptive typicality only, not also evaluative or all-things considered normality. Thus restricted, our constraint, in its official formulation, is necessary and sufficient for our truth conditions for counterfactuals together with the assumption that structural equations represent counterfactuals – a result that holds independently of the characterization of causal and backtracking counterfactuals. Furthermore, while I assume t to be a ranking function, my claims are meaningful also if t is a merely comparative pre-ordering that is reflexive and transitive, but not necessarily connected or antisymmetric.

9.6 Structural equations

It is time to deliver on my many promises. As mentioned in section 8.3, the conditions Lewis (1973a) imposes on overall similarity between possible worlds – and, hence, the logical properties of counterfactuals – do not distinguish between causal counterfactuals and other counterfactuals. This is attempted to be done only by Lewis (1979: 472)' "system of weights or priorities," an attempt that fails, as we have seen in section 8.3. It comes very close, though.

To characterize backtracking and causal counterfactuals, I will formulate four constraints[3] on extended acyclic causal models, restricted – as will always be the case from now on – to their purely descriptive elements. The first two constraints relate structural equations and typicality. The second characterizes backtracking counterfactuals by characterizing the causal structure all of which these hold fixed. Both can be motivated by Lewis (1979: 472)' conditions that

(1) [i]t is of the first importance to avoid big, widespread, diverse violations of law [... and that ...]

(3) [i]t is of the third importance to avoid even small, localized, simple violations of law.

However, there are differences. First, our first two constraints concern typicality and structural equations, not similarity and laws of nature. Second, our first two constraints are relative to an extended acyclic causal model, so not suited for a realist agenda without further assumptions on the adequacy of extended acyclic causal models. Third, while our first two constraints also appeal to violations – though of structural equations, not laws of nature – they are not formulated in entirely acausal terms. While we do not need to appeal to the difference between causal and backtracking counterfactuals, we do need to appeal to the difference between exogenous and endogenous variables. Fourth, our first two constraints do not appeal to the size of miracles, but whether they occur early or late in the causal hierarchy as it is pictured by the associated directed acyclic graph of an extended acyclic causal model. The second and third point will be true also of our third constraint which, additionally, concerns actuality. All five points will be true of our fourth constraint which characterizes causal counterfactuals.

[3] In stressing that it is an art to come up with an adequate model for a given scenario or case, Hitchcock (2007) states various constraints on the adequacy of a model (see also Hitchcock 2001 and Halpern & Hitchcock 2010). Hitchcock (2007)'s constraints concern the relation between the model and the case to be modeled. In contrast to these, our constraints are inherent to the model and independent of the case to be modeled.

Our constraints allow us to characterize, relative to a set of variables, both hard and soft interventions in the acausal terms of typicality and actuality, as well as the causal distinction between exogenous and endogenous variables. The latter is necessary for this characterization. This confirms the received interventionist view (Spirtes et al. 1993/2000, Pearl 2000/2009, Hitchcock & Woodward 2003, Woodward 2003; 2016, Woodward & Hitchcock 2003, and Glymour 2004) that causality cannot be analyzed or explicated in entirely acausal terms.

Just such a reduction is what Papineau (2022; ms)'s a posteriori analysis of the nature (rather than concept) of causation attempts. Papineau (2022; ms) assumes a bijective pairing of the endogenous and exogenous variables in a recursive set of equations – such as the structural equations of an acyclic causal model – as well as that the set of exogenous variables is probabilistically independent (see chapter 10 for probabilistic independence of sets of variables). This assumption implies that Pearl (2000/2009: 30)'s theorem 1.4.1 applies. As a result, the probability measure in the sense of which the set of exogenous variables is probabilistically independent satisfies the (causal) Markov condition for the directed acyclic graph that pictures the recursive set of equations (structural equations of the acyclic causal model): every variable is conditionally probabilistically independent of its non-descendents (non-effects) conditional on its parents (direct causes). The significance of this theorem is that it connects acyclic causal models and, hence, structural equations, to probability.

Unfortunately, Papineau (2022; ms)'s assumption presupposes the distinction between exogenous and endogenous variables. Treating the equations as mere equations without any structure and, hence, as symmetric, and not distinguishing between exogenous and endogenous variables renders this assumption insufficient for causal direction. To see this, consider Papineau (2022: 257)'s example relating schools S, parental income P, and examination results E:

$$P = e_P \tag{9.1}$$
$$S = a \times P + e_S \tag{9.2}$$
$$E = b \times P + c \times S + e_E \tag{9.3}$$

(a, b, and c are real numbers.) If $\{e_P, e_S, e_E\}$ is probabilistically independent and equations (9.1-9.3) are true, but were mere equations,

$$e_P = P \tag{9.4}$$
$$S = a \times e_P + e_S \tag{9.5}$$
$$E = b \times e_P + c \times S + e_E \tag{9.6}$$

would also be true and $\{P, e_S, e_E\}$ would also be probabilistically independent. In addition, if (9.1-9.3) is "expandable" (in the sense that "for any further variables correlated with those in [9.1-9.3], there is a larger system of equations covering those further variables that also satisfies exogenous independence and which has [9.1-9.3] as a subsystem" – Papineau 2022: 264), then so is (9.4-9.6). This means that the causal direction between P and e_P is not determined.

We start with some terminology relative to an extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. Say that possible world $w = \langle \vec{u}, v_1, \ldots, v_n \rangle$ *violates the structural equation* for the endogenous variable V_i, $1 \leq i \leq n$, if, and only if, $v_i \neq F_i(\vec{u}, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. Let $\mathcal{V}^*(w) \subseteq \mathcal{V}$ be the set of endogenous variables V_i such that w violates the structural equation for V_i. Next say that possible world w *weakly Halpern-dominates* possible world w' if, and only if, for each endogenous variable $X \in \mathcal{V}^*(w) \setminus \mathcal{V}^*(w')$ there is an endogenous variable $X' \in \mathcal{V}^*(w') \setminus \mathcal{V}^*(w)$ such that $X' \in An(X)$. Finally, say that possible world w *strongly Halpern-dominates* possible world w' if, and only if, w weakly Halpern-dominates w', but w' does not weakly Halpern-dominate w (so, $\mathcal{V}^*(w') \setminus \mathcal{V}^*(w)$ is not empty).[4]

Our first constraint says, among others, that a possible world that violates the structural equations for a set of endogenous variables is less typical than a possible world that violates the structural equations for only a proper subset of this set of endogenous variables. This is not all it says, though. The violation of the structural equation for an endogenous variable affects every variable that is – in the causal hierarchy as it is pictured by the associated directed acyclic graph – causally downstream of the former variable. For this reason the violation of the structural equation for an endogenous variable is worse – infinitely worse – than the violation of the structural equation for an endogenous variable that is causally downstream of the former variable. In Lewis (1986e: 55f)' terminology, the violation of the structural equation for an endogenous variable is a miracle that is infinitely bigger than the miracle that is the violation of the structural equation for an endogenous variable that is causally downstream of the former variable. This is why our first constraint has to be stated in terms of ancestors.

Here it is – except that the official formulation of our first constraint is *much* weaker than this paraphrase because it is restricted to possible worlds that agree on the value of every variable except for exactly one endogenous variable (the full version without this restriction is our second constraint):

[4]The terminology is to acknowledge the generous and helpful feedback I have received from Joseph Y. Halpern during his visits at Konstanz University in 2010 and 2011.

> **First Constraint** Extended acyclic causal models respect the causal structure. An extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *respects the causal structure* if, and only if, for all possible worlds w and w' in $\mathcal{W}$ that agree on the value of every variable except for exactly one endogenous variable: if w strongly Halpern-dominates w', then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$ in all contexts $\vec{u}$ in $R(\mathcal{U})$.

The idea is quite simple. First, associate with each possible world the set of endogenous variables whose structural equation this possible world violates. Then, when comparing two possible worlds for typicality, check whether they agree on the value of every variable except for exactly one endogenous variable. If so, ignore those endogenous variables whose structural equation is violated by both possible worlds. Finally, check whether, among the remaining endogenous variables, for each endogenous variable whose structural equation is violated by the first possible world there is an endogenous variable that is causally upstream of this variable and whose structural equation is violated by the second possible world. In addition, check whether the converse is not true. In other words, check if every violation by the first possible world is compensated for by a violation by the second possible world that concerns the structural equation for the same endogenous variable or one that is causally upstream of it. In addition, check if the converse is not true. If so, then the first possible world is less typical than the second possible world in every context. If the first possible world violates the structural equation for a set of endogenous variables that is a proper subset of the set of endogenous variables violated by the second possible world, we have the special case where, after ignoring the common violations, no violations by the first possible world are left. This is, of course, our constraint from the previous section, except that now it is restricted to possible worlds that agree on the value of every variable except for exactly one endogenous variable.

Woodward (2003: 141) can be read as endorsing our first constraint when he points to the following

> important general difference between Lewis's scheme and the manipulationist picture. On the manipulationist account, [...] "[l]ate" miracles, even numerous, are automatically preferred to "early" miracles, even if single. By contrast, in Lewis's theory, whether we [...] insert many late miracles [...] or whether instead we [insert some early miracle] [...] depends on whether [the effects] have many causes or just one. This sort of sensitivity leads to the insertion of miracles in what, intuitively, is the wrong place.

This is the sense in which an early miracle is infinitely bigger than a late miracle: the violation of the structural equation for any number of endogenous variables is compensated for by the violation of the structural equation for a single endogenous variable that is causally upstream of all the former variables.

Another difference is that, for Lewis (1979), a miracle involves two possible worlds: the possible world whose law of nature is violated and the possible world violating it. We can ignore the first possible world because the legal possible worlds are all governed by the same structural equations, while the illegal ones are governed by none and, in this sense, are lawless.

While our first constraint is formulated in terms of numerical rank-theoretic typicality functions, it is meaningful also if typicality is represented by a merely comparative pre-ordering, as in Halpern & Hitchcock (2015: 435) and Halpern (2016: 81). These authors implicitly subscribe to a slightly weaker version of our first constraint in these terms. The reason is that they identify the normality of a non-empty conjunction of atomic sentences in an extended directed acyclic causal model in a context with the normality of the unique possible world that is determined by the interventions making this conjunction true in this context. This possible world must not come out as less normal than any other possible world in which this conjunction is true in this context. Otherwise, a possible world can be strictly more normal than a sentence in which it is true, which I assume these authors want to avoid. In particular, this is so for conjunctions of atomic sentences that include every variable except for exactly one endogenous variable. The slightly weaker version of our first constraint in these terms follows by noting that the unique possible world for such a conjunction strongly Halpern-dominates every other possible world in which this conjunction is true in this context. The full version of our first constraint in these terms requires the unique possible world for conjunctions of atomic sentences that include every variable except for exactly one endogenous variable to come out as, not just not less normal, but strictly more normal than every other possible world in which this conjunction is true.

We continue with more terminology. $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ is a *typicality model* if, and only if, $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ is a signature and, for each possible world w in $\mathcal{W}$, ϱ_w is a ranking function on the power-set of $\mathcal{W}$. Rather than indexing the rank-theoretic typicality function to the context $\vec{u}$ or the legal possible world $w_{\vec{u}}$ that is determined by this context in extended acyclic causal models, the ranking function is now indexed to a possible world. The reason is that truth is a relation between sentences and possible worlds, not between sentences and contexts. This makes it necessary to be explicit about the exogenous variables. From now on $\mathcal{U}$ is the set of m exogenous variables $U_1, \ldots, U_m$, for some finite natural number m.

An atomic sentence $X_i = x$, $i = 1, \ldots, m + n$, is true in $w \in \mathcal{W}$ in $\mathcal{M}^*$ if, and only if, $w \in \{\langle u_1, \ldots, u_m, v_1, \ldots, v_m\rangle = \langle x_1, \ldots, x_{m+n}\rangle \in \mathcal{W} : x_i = x\}$. The truth conditions for negations, conjunctions, and disjunctions are given in the usual way. Where ϕ and ψ are arbitrary sentences, the default conditional $\phi \Rightarrow \psi$ is true in w in $\mathcal{M}^*$ if, and only if, ψ is true in the maximally ϱ_w-typical possible worlds in which ϕ is true. The counterfactual $\phi \mathrel{\square\!\!\rightarrow} \psi$ is true if, and only if, ψ is true in the maximally ϱ_w-typical possible worlds in which ϕ is true or – if w is less ϱ_w-typical than the maximally ϱ_w-typical possible worlds in which ϕ is true – ψ is true in all possible worlds in which ϕ is true and which are at least as ϱ_w-typical as w. For details, especially concerning infinitely atypical possible worlds, see sections 5.5 and 7.4.

Like our first constraint, these truth conditions for default conditionals and counterfactuals are meaningful also if typicality is represented by a pre-ordering. For default conditionals we can follow Halpern (2003/2017: 304ff): we identify the maximally typical possible worlds in which a sentence is true with the possible worlds in which the sentence is true for which there is no possible world in which the sentence is true and which is more typical. For counterfactuals we cannot follow Halpern (2003/2017: 319f), but proceed as follows: if the actual world is less typical than a (maximally typical) possible world in which a sentence is true, we identify the possible worlds in which a sentence is true which are at least as typical as the actual world with the possible worlds in which the sentence is true which are not less typical than the actual world. There is one caveat, though. Typicality is assumed to satisfy the limit assumption in the sense that for each pre-ordering $\preceq_w$ of atypicality in possible world w and each consistent proposition $A \subseteq \mathcal{W}$ there is a possible world w' in A such that for all possible worlds w'' in A: w' is not less typical in w than w'' ($w'' \not\prec_w w'$, that is, $w'' \not\preceq_w w'$ or $w' \preceq_w w''$). For the same reason as before, Halpern & Hitchcock (2015) and Halpern (2016) implicitly subscribe to this assumption restricted to (propositions that are expressed by) non-empty conjunctions of atomic sentences, which are the only sentences that have typicality values in the framework of these authors.

In an extended acyclic causal model the structural equations are given and then used to define truth conditions for a limited set of counterfactuals. In a typicality model the default conditionals and counterfactuals are given via the rank-theoretic typicality functions ϱ_w (plus their truth conditions). Therefore, we have to say what it means for a structural equation represented by some function F to hold in a typicality model. This can be done in more than one way which reflects the incompleteness of the modal component $\mathcal{F}$ of a causal model $\langle \mathcal{S}, \mathcal{F}\rangle$ as compared to the modal component $(\varrho_w)_{w\in\mathcal{W}}$ of a typicality model $\langle \mathcal{S}, (\varrho_w)_{w\in\mathcal{W}}\rangle$.

Without claiming this to be the best way to reproduce structural equations, I will proceed in a way that appeals to nothing causal but the distinction between exogenous and endogenous variables. This way employs default conditionals and renders structural equations absolutely necessary in the sense that they hold in every possible world in a typicality model. These include the possible worlds that, in the extended acyclic causal model, are illegal in the sense of violating at least one structural equation. For this reason we cannot use counterfactuals or other conditionals that validate modus ponens. Instead, we need to use conditionals such as default conditionals that allow for exceptions. Once we have reproduced structural equations and, hence, the distinction between legal and illegal possible worlds in this round-about way, we can then characterize them also in terms of counterfactuals that hold in all legal possible worlds.

An alternative, more direct way to reproduce structural equations is to employ counterfactuals, but define truth to be a relation between sentences and contexts rather than sentences and possible worlds. This way, too, appeals to nothing causal but the distinction between exogenous and endogenous variables. Specifically, the distinction between causal and backtracking counterfactuals is not required for this way of reproducing structural equations.

The second way may be preferred by proponents of an alternative semantics for counterfactuals because it allows them to do without typicality. For us it is the other way round, as we show the second way to be viable by showing the first way to be viable. Another reason for proceeding in the first way is that it suggests how interventionist counterfactuals can be generalized to causal counterfactuals that allow for soft interventions as antecedents. Once structural equations have been characterized in terms of counterfactuals that are true in all legal possible worlds, we can generalize interventionist counterfactuals to causal counterfactuals by dropping an assumption: we let each possible world have its own, possibly empty set of structural equations, including possible worlds that previously were not only illegal, but lawless in the sense of having no structural equations. This renders structural equations "absolutely contingent": no two possible worlds need to have the same, and no possible world needs to have any, structural equations. In particular, unlike in causal models, in typicality models a possible world can fail to violate every structural equation in a given set of such without being governed by their collection. That is, unlike causal models, typicality models do not imply that there are no "accidental generalizations" (or rather, material conditionals). This means we have use for absolutely necessary default conditionals and absolutely contingent (backtracking or causal) counterfactuals, but not contingent laws of nature (section 7.3). Occam's razor implies that these do not exist.

$$T \longleftrightarrow S$$

Figure 9.10: boat (without the restriction to endogenous variables)

We need a bit more terminology. Say that F is *eligible* for V_i if, and only if, F is a function from $\mathcal{W}_i = \times_{X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_i\}} R(X)$ into $R(V_i)$, where V_i is an endogenous variable. It is here that we appeal to the causal distinction between exogenous and endogenous variables: there are no eligible functions for exogenous variables. A function $F : \mathcal{W}_i \to R(V_i)$ which is eligible for V_i *holds* in a typicality model $\mathcal{M}^*$ if, and only if, for every possible world w in $\mathcal{W}$, the following default conditionals are all true in w in $\mathcal{M}^*$: where $\vec{w}_i$ is in $\mathcal{W}_i$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \Rightarrow V_i = F_i(\vec{w}_i).$$

For an eligible function to hold in a typicality model all these default conditionals must be true, and they must be true in all possible worlds. In contrast to default conditionals in general, whose truth value is world-dependent, structural equations hold world-independently. They are absolutely necessary, but allow for exceptions (see Cartwright 1980 on exceptions of laws of nature, as well as the literature on ceteris paribus laws: for instance, *Erkenntnis* **57** (3), **79** (10)).

Theorem 9. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure if, and only if, there is a typicality model $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ with the same signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ such that:*

T for every context $\vec{u}$ in $R(\mathcal{U})$: $\varrho_{\vec{u}} = \varrho_{w_{\vec{u}}}$, where $w_{\vec{u}}$ is the unique legal world determined by $\vec{u}$ in $\mathcal{M}$; and

SE $F_i \in \mathcal{F}$ if, and only if, F_i holds in $\mathcal{M}^*$ if, and only if, for all $\vec{w}_i$ in $\mathcal{W}_i$: $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \mathrel{\square\!\!\rightarrow} V_i = F_i(\vec{w}_i)$ is true in all legal possible worlds $w_{\vec{u}}$ in $\mathcal{M}^*$.

PROOF: See the appendix to chapter 9. *Q.E.D.*

Our first result covers context-independent typicality as special case and holds also if typicality is represented by a pre-ordering. As the boat example shows, it requires the causal distinction between exogenous and endogenous variables, something Huber (2013) fails to stress. Without restricting eligible functions to endogenous variables, we get not only the function $F_S : t \mapsto 1 - t$ describing the structural equation $S = \neg T$, but also the function $F_T : s \mapsto 1 - s$ describing the structural equation $T = \neg S$. The resulting extended cyclic causal model is pictured by a directed cyclic graph (figure 9.10).

9.7 Backtracking and causal counterfactuals

Backtracking counterfactuals hold fixed all of a certain causal structure. Causal counterfactuals hold fixed what remains of this causal structure after holding fixed what is not causally downstream of the antecedent. In a causal model, this causal structure is given by the structural equations represented by a set of functions $\mathcal{F}$. This causal structure is the same for every legal possible world, and non-existent for illegal possible worlds that are better characterized as lawless. In a typicality model, this causal structure is given by the typicality function ϱ_w of possible world w, as well as the causal distinction between exogenous and endogenous variables. This causal structure need not be the same for any two possible worlds, and may fail to exist for every possible world.

In a typicality model, the structural equation for endogenous variable V_i that is represented by F_i holds at possible world w if, and only if, all of the following counterfactuals are true at w: where $\vec{w}_i$ is in $\mathcal{W}_i$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \,\square\!\!\rightarrow\, V_i = F_i(\vec{w}_i)\,.$$

A possible world is not governed by a structural equation for V_i if, and only if, for every function F that is eligible for V_i there is at least one $\vec{w}_i$ in $\mathcal{W}_i$ such that the counterfactual $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \,\square\!\!\rightarrow\, V_i = F(\vec{w}_i)$ is false at w.

There are typicality models with possible worlds that are not governed by any structural equation, but at which non-trivial counterfactuals are true. We will see in the next section that these counterfactuals are enough for these possible worlds to have non-trivial "causal laws" which determine a non-trivial causal structure that can be pictured by a directed acyclic graph. An alternative way to get such causal structure from typicality in the absense of structural equations rests on the following result. A directed acyclic graph whose nodes, i.e., variables, are linearly ordered (so that the parents of any variable, i.e., the variables that are directly causally relevant to this variable, precede the latter in this linear ordering) is determined by the conditional independence relation of every rank-theoretic typicality function satisfying the Markov and minimality conditions for this graph in their rank-theoretic formulations (Spohn 2012: ch. 7). The same is true also if typicality is represented by a pre-ordering, as shown by Halpern & Hitchcock (2013: sct. 3). The assumption that the variables are linearly ordered or some alternative assumption is necessary for this result. This assumption implies that the very first variable in the linear ordering is exogenous. Other than that, this assumption cannot be compared to our causal assumption that specifies which variables are exogenous and which endogenous.

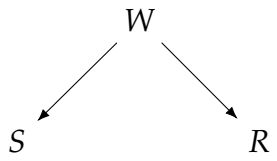

Figure 9.11: Ida doing her thing

The former assumption is generally motivated by the idea that actual causes temporally precede their effects (section 8.2). This means that the linear ordering is a temporal one and that we talk about efficient causation between events or related relata that are formally represented by singular variables. Our assumption is more flexible. It applies to other notions of causation besides efficient causation and exogenous variables need not be temporally prior to enodgenous ones (for instance, they can be model-relative unmoved movers; cf. Aristoteles BCE/1984).

Consider again the example from section 8.3, this time in the language of causal models. Let exogenous variable W take on value 1 if Ida has wine the night before, and 0 otherwise. Let endogenous variable S take on value 1 if Ida sleeps in, and 0 otherwise. Let endogenous variable R take on value 1 if Ida goes for a run in the morning, and 0 otherwise. The functions $F_S : w \mapsto w$ and $F_R : w \mapsto w$ describe the following structural equations which are pictured by a directed acyclic graph (figure 9.11):

$$S = W$$

$$R = W$$

Suppose the corresponding extended directed acyclic causal model respects the causal structure and, in the actual context, Ida has wine the night before (so, sleeps in and goes for a run in the morning). The typicality models with the same signature that exist for this extended acyclic causal model according to theorem 9 all have the same structural equations, as well as the same typicality functions in all legal possible worlds, including the actual world @. The typicality function $\varrho_@$ for the latter is such that the following counterfactuals are all true in @:

$$\begin{aligned}
&W = 1 \wedge R = 1 \boxdot\!\!\rightarrow S = 1, W = 1 \wedge R = 0 \boxdot\!\!\rightarrow S = 1,\\
&W = 0 \wedge R = 1 \boxdot\!\!\rightarrow S = 0, W = 0 \wedge R = 0 \boxdot\!\!\rightarrow S = 0,\\
&W = 1 \wedge S = 1 \boxdot\!\!\rightarrow R = 1, W = 1 \wedge S = 0 \boxdot\!\!\rightarrow R = 1,\\
&W = 0 \wedge S = 1 \boxdot\!\!\rightarrow R = 0, W = 0 \wedge S = 0 \boxdot\!\!\rightarrow R = 0.
\end{aligned}$$

The same is true for the interventionist counterfactuals in the extended acyclic causal model in the actual context:

$$W = 1 \wedge R = 1 \mathrel{-\!\!\!\exists} S = 1, W = 1 \wedge R = 0 \mathrel{-\!\!\!\exists} S = 1,$$
$$W = 0 \wedge R = 1 \mathrel{-\!\!\!\exists} S = 0, W = 0 \wedge R = 0 \mathrel{-\!\!\!\exists} S = 0,$$
$$W = 1 \wedge S = 1 \mathrel{-\!\!\!\exists} R = 1, W = 1 \wedge S = 0 \mathrel{-\!\!\!\exists} R = 1,$$
$$W = 0 \wedge S = 1 \mathrel{-\!\!\!\exists} R = 0, W = 0 \wedge S = 0 \mathrel{-\!\!\!\exists} R = 0.$$

So far, so good. Now consider:

BC If Ida had not slept in, she *must* not have had [*that* would have been *because* she did not have] wine the night before, and, *so*, she would not have gone for a run in the morning.

CC *Even* if Ida had not slept in, she would *still* have had wine the night before, and she would *still* have gone for a run in the morning.

The backtracking counterfactual BC holds fixed the causal structure in its entirety. By contrast, the causal counterfactual CC holds fixed the causal structure to the extent possible. In the extended acyclic causal model, the first of the following two interventionist counterfactuals is false in the actual context, while the second is true:

$$S = 0 \mathrel{-\!\!\!\exists} W = 0 \wedge R = 0,$$
$$S = 0 \mathrel{-\!\!\!\exists} W = 1 \wedge R = 1.$$

This is, of course, as it should be because interventionist counterfactuals are (a special case of) causal counterfactuals.

What about the counterfactuals in the typicality models? This depends on whether we understand them as backtracking counterfactuals whose connective will be symbolized by '$\bigcirc\!\!\rightarrow$' or as causal counterfactuals whose connective will be symbolized by '$\boxdot\!\!\rightarrow$.' So far, we have appealed only to what backtracking and causal counterfactuals have in common: $\Box\!\!\rightarrow$ can be understood as either a third connective that is restricted to what causal and backtracking counterfactuals have in common, or else ambiguous between the two connectives $\bigcirc\!\!\rightarrow$ and $\boxdot\!\!\rightarrow$ when all one is relying on is what the two have in common. Among others, this means that the structural equation for endogenous variable V_i that is represented by F_i holds at possible world w in a typicality model $\mathcal{M}^*$ if, and only if, all of the following backtracking counterfactuals are true at w: where $\vec{w}_i$ is in $\mathcal{W}_i$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \mathrel{\bigcirc\!\!\rightarrow} V_i = F_i(\vec{w}_i).$$

This in turn is the case if, and only if, all of the following causal counterfactuals are true at w: where $\vec{w}_i$ is in $\mathcal{W}_i$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \boxdot\!\!\rightarrow V_i = F_i(\vec{w}_i)\,.$$

Only if we consider counterfactuals whose antecedents do not specify the value of every variable except for exactly one endogenous variable do backtracking and causal counterfactuals come apart.

Let us return to our example's extended acyclic causal model. In the typicality models that reproduce it according to theorem 9, the first of the following two backtracking counterfactuals must come out as true at the actual world, and the second must come out as false:

$$S = 0 \bigcirc\!\!\rightarrow W = 0 \wedge R = 0,$$
$$S = 0 \bigcirc\!\!\rightarrow W = 1 \wedge R = 1.$$

By contrast, the first of the following two causal counterfactuals must come out as false at the actual world, and the second must come out as true:

$$S = 0 \boxdot\!\!\rightarrow W = 0 \wedge R = 0,$$
$$S = 0 \boxdot\!\!\rightarrow W = 1 \wedge R = 1.$$

Our first constraint characterizes what causal and backtracking counterfactuals have in common. This is more than might appear, as our first constraint is stronger than I have previously led on.

Theorem 10. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure if, and only if, for all possible worlds w and w' in $\mathcal{W}$ that agree on the value of every exogenous variable: if w strongly Halpern-dominates w', then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$ in all contexts $\vec{u}$ in $R(\mathcal{U})$.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

Our first constraint says that the causal structure within each context is held fixed at all contexts. Both backtracking and causal counterfactuals hold fixed this much causal structure. As theorem 9 shows, it is enough to reproduce structural equations in typicality models. Backtracking and causal counterfactuals differ in what they hold fixed across contexts. Backtracking counterfactuals hold fixed, in any given context, also the cross-contextual causal structure, even if this comes at the cost of changing the context itself. By contrast, causal counterfactuals hold fixed, in any given context, what is actually the case in this context, even if this comes at the cost of changing the cross-contextual causal structure. Let us make these informal characterizations precise.

> **Characteristic Constraint for Backtracking Counterfactuals ($\bigcirc\!\!\rightarrow$)** Extended acyclic causal models respect the whole causal structure, and nothing but the causal structure. An extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *respects the whole causal structure, and nothing but the causal structure* if, and only if, for all possible worlds w and w' in $\mathcal{W}$: if w strongly Halpern-dominates w', then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$ in all contexts $\vec{u}$ in $R(\mathcal{U})$.

Our second constraint is a strengthening of our first constraint. It says that strong Halpern-dominance guides typicality in every context not just for possible worlds that agree on the value of every variable except for exactly one endogenous variable – that is, according to theorem 10, possible worlds that agree on the values of all exogenous variables – but all possible worlds. If the extended acyclic causal model of our example respects the whole causal structure, and nothing but the causal structure, the typicality models reproducing it according to theorem 9 render BC true and CC false. That said, it is nothing but an assumption of mine that our second constraint characterizes backtracking counterfactuals.

To characterize causal counterfactuals, I will formulate two more constraints. Our third constraint relates "typicality" and actuality and can be motivated by Lewis (1979: 472)' condition that

(2) [i]t is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

However, as with our first constraint, there are differences. Our third constraint too is relative to an extended acyclic causal model and appeals to the causal distinction between exogenous and endogenous variables. In addition, it concerns agreement in the values of exogenous variables in the actual world, not perfect match (or approximate similarity) of particular actual-world-fact.

As with our first constraint, we start with some terminology relative to an extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. Say that possible world $w = \langle u_1, \ldots, u_m, \vec{v} \rangle$ *differs* from possible world $w^+ = \langle u_1^+, \ldots, u_m^+, \vec{v}^+ \rangle$ in the value for the exogenous variable U_i if, and only if, $u_i \neq u_i^+$. Let $\mathcal{U}^*_{w^+}(w)$ be the set of exogenous variables such that w differs from w^+ in their value. Next say that possible world w *weakly dominates* possible world w' in terms of *focus on possible world* w^+ if, and only if, $\mathcal{U}^*_{w^+}(w) \subseteq \mathcal{U}^*_{w^+}(w')$. Finally, say that possible world w *strongly dominates* possible world w' in terms of *focus on possible world* w^+ if, and only if, w weakly dominates w' in terms of focus on w^+, but w' does not weakly dominate w in terms of focus on w^+ (so, $\mathcal{U}^*_{w^+}(w') \setminus \mathcal{U}^*_{w^+}(w)$ is not empty).

Our third constraint says that a possible world that differs from the actual world in the value for a set of exogenous variables is less "typical" in the actual world than a possible world that differs from the actual world in the value for only a proper subset of this set of exogenous variables. This is not how to think of it, though. Our third constraint is not intended to be about typicality, as our first and second constraint are, but about what is held fixed in the evaluation of a causal counterfactual. For this reason I will present it in a different light shortly.

> **Third Constraint** Extended acyclic causal models are focused on actuality. An extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *is focused on actuality* if, and only if, for all contexts $\vec{u}$ in $R(\mathcal{U})$ and all possible worlds w and w' in $\mathcal{W}$: if w strongly dominates w' in terms of focus on $w_{\vec{u}}$, then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$.

The idea is quite simple. First, associate with each possible world the set of exogenous variables whose value is different in the actual world. Then, when comparing two possible worlds for "typicality" in the actual world, check whether each exogenous variable on whose value the first possible world disagrees with the actual world is also an exogenous variable on whose value the second possible world disagrees with the actual world. In addition, check if the converse is not true. If so, the first possible world is more "typical" in the actual world than the second possible world.

Our first and second constraint are global: if they rule that a possible world is more typical than another in one context, then this is so in all contexts. Our third constraint is local: it concerns "typicality" in the actual context. This is why we now quantify over contexts at the beginning of the relevant clause, whereas previously we quantified over contexts at the end of the relevant clause. Our fourth constraint is a mixture of global and local.

> **Characteristic Constraint for Causal Counterfactuals ($\boxdot\!\!\rightarrow$)** Extended acyclic causal models respect the causal structure and are focused on actuality.

Like our second, our fourth constraint is a strengthening of our first constraint. It adds our third constraint to our first constraint as a conjunct. If the extended acyclic causal model of our example respects the causal structure and is focused on actuality, the typicality models reproducing it according to theorem 9 render CC true and BC false. Furthermore, given our truth conditions for counterfactuals, interventionist counterfactuals are a special case of causal counterfactuals if, and only if, our fourth constraint characterizes causal counterfactuals.

Theorem 11. *Extended acyclic causal model* $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *respects the causal structure and is focused on actuality if, and only if, there is a typicality model* $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ *with the same signature* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ *such that:*

T for every context $\vec{u}$ *in* $R(\mathcal{U})$*:* $\varrho_{\vec{u}} = \varrho_{w_{\vec{u}}}$*, where* $w_{\vec{u}}$ *is the unique legal world determined by* $\vec{u}$ *in* $\mathcal{M}$*; and*

C for every sentence ϕ *in the language of the generalized version of Halpern & Hitchcock (2010) and context* $\vec{u}$ *in* $R(\mathcal{U})$*:*
ϕ *is true in* $\mathcal{M}$ *in* $\vec{u}$ *if, and only if,* ϕ *is true in* $w_{\vec{u}}$ *in* $\mathcal{M}^*$*.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

For ease of readability, clause C identifies the interventionist counterfactual $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \mathrel{-\!\!\!\exists} \phi$ whose truth conditions are stated in terms of causal models and contexts with the causal counterfactual $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \boxdot\!\!\rightarrow \phi$ whose truth conditions are stated in terms of possible worlds and typicality models. Specifically, C says, among other things, that, for any $k \geq 1$ distinct variables $X_1, \ldots, X_k$ (with possible values $x_1, \ldots, x_k$, respectively) and Boolean combination of atomic sentences ϕ: $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \mathrel{-\!\!\!\exists} \phi$ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \boxdot\!\!\rightarrow \phi$ is true in $w_{\vec{u}}$ in $\mathcal{M}^*$.

Our third and fourth constraint are weaker than I have previously led on.

Theorem 12. *If extended acyclic causal model* $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *respects the causal structure, then* $\mathcal{M}$ *is focused on actuality if, and only if, for all contexts* $\vec{u}$ *in* $R(\mathcal{U})$ *and possible worlds* w *and* w' *in* $\mathcal{W}$ *that agree on the value of every variable except for exactly one exogenous variable: if* w *strongly dominates* w' *in terms of focus on* $w_{\vec{u}}$*, then* $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$*.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

Our third and fourth constraint are meaningful also if typicality is represented by a pre-ordering. Theorems 10-12 hold also in this case. The former covers context-independent typicality as a special case. The latter two are trivialized by it, which is why they will be presented in a different light shortly. Before doing so, note that our fourth constraint delivers the revision of Lewis (1979)' system promised at the end of section 8.3. It renders perfect match of certain – viz., exogenous – particular fact as important as avoiding certain – viz., intra-contextual – miracles, while perfect match of other particular fact and avoidance of other miracles do not matter at all. The relative importance of miracles within a context is specified in terms of a lexicographic order. And, it can be characterized in merely comparative, albeit neither model-independent nor entirely acausal terms.

Our third and fourth constraint tie typicality to the values of the exogenous variables in the actual context or possible world. Typicality and actuality can come apart, though. Our third and fourth constraint are not intended to deny this, but to specify what is held fixed in the evaluation of causal counterfactuals. The idea is that, in evaluating causal counterfactuals, we do not use non-conditional typicality, but a form of conditional typicality that holds fixed certain aspects of actuality.

Consider an extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$. Suppose we allow only endogenous variables in the antecedents of causal counterfactuals and all rank-theoretic typicality functions $\varrho_{\vec{u}}$ are regular. In this case one way to define the *contextualization* of $\mathcal{M}$ is to replace the non-conditional typicality functions $\varrho_{\vec{u}}(\cdot)$ by the conditional typicality functions $\varrho_{\vec{u}}\left(\cdot \mid \vec{\mathcal{U}} = \vec{u}\right)$. The latter hold fixed the values of the exogenous variables at their actual values.

The contextualization of an extended acyclic causal model can be defined in a different way if we consider the language of the generalized version of Hitchcock & Halpern (2010) which allows for exogenous variables in the antecedents of causal counterfactuals, if not all rank-theoretic typicality functions are regular, or if typicality is represented by a pre-ordering (see the appendix for details). Either way, the following holds.

Theorem 13. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure if, and only if, the contextualization of $\mathcal{M}$ respects the causal structure and is focused on actuality.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

Contextualization is not the only operation with this property. Put differently, theorem 13 holds for different definitions of contextualization. That said, there is no operation for which theorem 13 holds that changes the ordering of typicality among two possible worlds in a context $\vec{u}$ only if these two possible worlds are such that one of them strongly dominates the other in terms of focus on the legal possible world $w_{\vec{u}}$ (so that the dominating possible world must come out as more typical in this context than the dominated possible world in order for the resulting extended acyclic causal model to be focused on actuality; again, see the appendix for details).

To conclude this section, note that extended acyclic causal models that respect the whole causal structure, and nothing but the causal structure which characterize backtracking counterfactuals are a special case of extended acyclic causal models that respect the causal structure. Contextualization and theorem 13 also apply to them.

9.8 Causality

We already have relativized typicality and structural equations to possible worlds, but, so far, kept the assumption that the causal distinction between exogenous and endogenous variables is the same for all possible worlds. We also have not yet characterized backtracking or causal counterfactuals if there is no structural equation for some endogenous variable at a possible world. Time to do better.

More terminology. $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ is a *causality model* if, and only if, $\mathcal{X}$ is a finite set of variables, $R : \mathcal{X} \to \mathcal{R}$ assigns to each variable $X \in \mathcal{X}$ its range $R(X) \subseteq \mathcal{R}$, and, for each possible world w in $\mathcal{W} = \times_{X\in\mathcal{X}} R(X)$, e_w divides $\mathcal{X}$ into a set of definitely exogenous variables $\mathcal{U}_w \subseteq \mathcal{X}$ and a set of potentially endogenous variables $\mathcal{V}_w = \mathcal{X} \setminus \mathcal{U}_w$, while ϱ_w is a ranking function on the power-set of $\mathcal{W}$. $\langle \mathcal{X}, R\rangle$ is a *specification of the variables*. $\mathcal{V}_w$ is the set of *potentially* endogenous variables because it can happen that a variable in it fails to have a parent (to be defined momentarily) in which case it is exogenous. By contrast, $\mathcal{U}_w$ contains only variables that are definitely exogenous, even if the ranking function ϱ_w would specify parents for a variable in it if this variable were not banned from having parents by stipulation. Alternatively, we can stipulate that, in each possible world w, e_w and ϱ_w are *aligned* so that every variable in $\mathcal{V}_w$ has at least one parent. Obviously, every typicality model is a causality model, but not conversely.

Since it is possible that there is no structural equation for some endogenous variable in a possible world, we need to generalize this concept. In a causality model, the *causal law* for potentially endogenous variable V_i at possible world w is the collection of the following counterfactuals that are true in w: where $\vec{w}_i$ is in $\mathcal{W}_i = \times_{X\in\mathcal{X}\setminus\{V_i\}} R(X)$ and R_i is any subset of $R(V_i)$,

$$\mathcal{X} \setminus \vec{\{V_i\}} = \vec{w}_i \,\square\!\!\rightarrow\, V_i \in R_i.$$

(The sentence $X_j \in R_j$, for $R_j \subseteq R\left(X_j\right)$, is true in $w \in \mathcal{W}$ in C if, and only if, $w \in \left\{\langle x_1, \ldots, x_{m+n}\rangle \in \mathcal{W} : x_j \in R_j\right\}$. It can be construed as atomic sentence or, if one does not mind infinitely long sentences, as the disjunction of atomic sentences $X_j = x_j$, for all $x_j \in R_j$. I will continue to focus on semantic considerations and bracket these syntactic issues, except for mentioning that $X_j \in R\left(X_j\right)$ and $\neg\left(X_j \in R_j^1 \wedge X_j \in R_j^2\right)$ for disjoint R_j^1 and R_j^2 must be theorems.) Obviously, every function F_i representing the structural equation for endogenous variable V_i in a(n acyclic) causal model is a causal law in a possible world in a(n acyclic – to be defined momentarily) causality model, but not conversely.

Possible world w' *violates* the causal law for potentially endogenous variable V_i at possible world w if, and only if, for some subset R_i of $R(V_i)$,

$$\mathcal{X} \setminus \vec{\{}V_i\} = \vec{w_i'} \boxdot\!\!\rightarrow V_i \in R_i$$

is true in w and $V_i \in R_i$ is false in w', where $\vec{w_i'}$ are the values of $\mathcal{X} \setminus \vec{\{}V_i\}$ in w'.

Besides violation of a causal law, the other concept we need is that of ancestor of a potentially endogenous variable. Variable V_i that is potentially endogenous at possible world w is *co-determined* in w by variable Y if, and only if, there are $\vec{w_i}$ and $\vec{w_i}'$ in $\mathcal{W}_i$ that differ only in the value from $R(Y)$ such that, for some subset R_i of $R(V_i)$, one of the following sentences is true at w and the other is not:

$$\mathcal{X} \setminus \vec{\{}V_i\} = \vec{w_i} \boxdot\!\!\rightarrow V_i \in R_i \qquad \mathcal{X} \setminus \vec{\{}V_i\} = \vec{w_i}' \boxdot\!\!\rightarrow V_i \in R_i$$

One important caveat: if typicality is represented by a pre-ordering that is not connected, these definitions are restricted to subsets R_i of $R(V_i)$ that contain exactly one element. This means that we are quantifying over elements v_i in – rather than subsets R_i of – $R(V_i)$. In this case the causal law for V_i may be empty.

A variable V_i that is potentially endogenous in possible world w may not be co-determined in w by any variable Y even if the causal law for V_i in w is non-trivial (as well as violated by another possible world). For instance, this happens if there is a v_i in $R(V_i)$ such that for all $\vec{w_i}$ in $\mathcal{W}_i$: $\mathcal{X} \setminus \vec{\{}V_i\} = \vec{w_i} \boxdot\!\!\rightarrow V_i = v_i$ is true at w. If readers prefer, they can rule out this case by fiat and stipulate that e_w and ϱ_w are aligned.

In a causality model, the set of parents of variable X in possible world w, $Pa_w(X)$, is the set of variables Y such that X is co-determined in w by Y. In a causality model, a possible world w is acyclic if, and only if, it is not the case that there are m variables $X_1, \ldots, X_m$, for some natural number $m \geq 2$, such that, in w, X_{j+1} is co-determined by X_j for $j = 1, \ldots, m-1$, and X_1 is co-determined by X_m. A causality model is acyclic if, and only if, every possible world in it is acyclic. Every acyclic possible world in a causality model can be pictured by a directed acyclic graph with exactly one node for each variable that specifies the value of this variable in this possible world, and with arrows into each variable from all and only its parents in this possible world. Since causal laws generalize structural equations and each directed acyclic graph pictures the structural equations of some causal model, we get the following result, where $\langle\langle \mathcal{X}, R\rangle, \rightarrow\rangle$ is a directed acyclic graph if, and only if, $\langle \mathcal{X}, R\rangle$ is a specification of the variables, $\rightarrow \subseteq \mathcal{X} \times \mathcal{X}$, and it is not the case that there are m variables $X_1, \ldots, X_m$, for some natural number $m \geq 2$, such that, $X_j \rightarrow X_{j+1}$ (that is, $\langle X_j, X_{j+1}\rangle \in \rightarrow$) for $j = 1, \ldots, m-1$, and $X_m \rightarrow X_1$. This result holds also if typicality is represented by a pre-ordering.

Theorem 14. *For each directed acyclic graph* $\langle\langle \mathcal{X}, R\rangle, \rightarrow\rangle$ *there is at least, but not necessarily exactly, one possible world* w^+ *in an acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *such that for all variables* X *and* Y *in* $\mathcal{X}$*:* $X \rightarrow Y$ *if, and only if,* X *co-determines* Y *in* w^+*. For each acyclic possible world* w^+ *in a causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *there is exactly one directed acyclic graph* $\langle\langle \mathcal{X}, R\rangle, \rightarrow\rangle$ *such that for all variables* X *and* Y *in* $\mathcal{X}$*:* $X \rightarrow Y$ *if, and only if,* X *co-determines* Y *in* w^+*.*

PROOF: Follows from the preceeding remarks. *Q.E.D.*

A variable X that is potentially endogenous in possible world w is (actually) endogenous in w if, and only if, X has at least one parent in w. A variable X is (actually) exogenous in possible world w if, and only if, X is definitely exogenous in w or potentially, but not actually endogenous in w. For future reference, call a causality model *structured through and through* if, and only if, for every possible world w, every variable V_i that is potentially endogenous in w, and every $\vec{w_i}$ in $\mathcal{W}_i = \times_{X\in\mathcal{X}\setminus\{V_i\}} R(X)$ there is a v_i in $R(V_i)$ such that the following counterfactual is true in w:

$$\mathcal{X} \setminus \vec{\{V_i\}} = \vec{w_i} \boxdot\!\!\rightarrow V_i = v_i$$

With the concepts of violation of a causal law and ancestor of a variable, we can reproduce the concept of Halpern-dominance for an acyclic possible world in a causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$. Let $\mathcal{V}^*_{w^+}(w)$ be the set of variables V_i that are (actually) endogenous in w^+ and such that w violates the causal law for V_i at w^+. Say that possible world w *weakly dominates* possible world w' in possible world w^+ if, and only if, for each $X \in \mathcal{V}^*_{w^+}(w) \setminus \mathcal{V}^*_{w^+}(w')$ there is an $X' \in \mathcal{V}^*_{w^+}(w') \setminus \mathcal{V}^*_{w^+}(w)$ such that $X' \in An_{w^+}(X)$. Next say that possible world w *strongly dominates* possible world w' in possible world w^+ if, and only if, w weakly dominates w' in w^+, but w' does not weakly dominate w in w^+(so, $\mathcal{V}^*_{w^+}(w') \setminus \mathcal{V}^*_{w^+}(w)$ is not empty). Finally, with the understanding that the exogenous variables now are the variables that are (actually) exogenous in w^+, we continue to say that possible world w *strongly dominates* possible world w' in terms of *focus on possible world* w^+ if, and only if, w weakly dominates w' in terms of focus on w^+, but w' does not weakly dominate w in terms of focus on w^+.

Now we can formulate our second and fourth constraint which characterize, respectively, backtracking and causal counterfactuals for acyclic causality models (that satisfy our first constraint for acyclic causality models) rather than extended acyclic causal models. Both are meaningful also if typicality is represented by a pre-ordering.

Characteristic Constraint for Backtracking Counterfactuals in Acyclic Causality Models ($\bigcirc\!\!\rightarrow$) Acyclic causality models respect the whole causal structure, and nothing but the causal structure. An acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the whole causal structure, and nothing but the causal structure* if, and only if, for all possible worlds w^+, w, and w' in $\mathcal{W}$: if w strongly dominates w' in w^+, then $\varrho_{w^+}(w) < \varrho_{w^+}(w')$.

Characteristic Constraint for Causal Counterfactuals in Acyclic Causality Models ($\boxdot\!\!\rightarrow$) Acyclic causality models respect the causal structure and are focused on actuality. An acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure and is focused on actuality* if, and only if, for all possible worlds w^+, w, and w' in $\mathcal{W}$: if w and w' agree on the value of all variables except for exactly one variable that is (actually) endogenous in w^+ and w strongly dominates w' in w^+, then $\varrho_{w^+}(w) < \varrho_{w^+}(w')$; and if w strongly dominates w' in terms of focus on w^+, then $\varrho_{w^+}(w) < \varrho_{w^+}(w')$.

Both constraints are local rather than global because, unlike structural equations in causal models, causal laws in causality models are as world-dependent as Lewis (1979)' laws of nature. For this reason the latter constraint delivers a revision of Lewis (1979)' system that is even closer to the original than the revision delivered by our fourth constraint in the previous section.

Our second and fourth constraint are incompatible provided there are at least two contexts. In general, so are their formulations for acyclic causality models. In section 8.3 I noted that I think of backtracking and causal counterfactuals as employing different connectives that have many features in common, including our first constraint, but differ in other respects. Let us formulate the former for acyclic causality models.

First Constraint for Acyclic Causality Models Acyclic causality models respect the causal structure. An acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure* if, and only if, for all possible worlds w^+, w, and w' in $\mathcal{W}$: if w and w' agree on the value of all variables except for exactly one variable that is (actually) endogenous in w^+ and w strongly dominates w' in w^+, then $\varrho_{w^+}(w) < \varrho_{w^+}(w')$.

One of the respects in which they differ is the centering axiom 12. In the legal possible worlds of the typicality models which reproduce extended acyclic causal models that respect the causal structure according to theorem 9, it holds for causal, but not backtracking counterfactuals (even if we restrict ourselves to the language of the generalized version of Halpern & Hitchcock 2010). If we restrict ourselves to the language of the generalized version of Halpern & Hitchcock (2010), the same is true for the law of conditional excluded middle 0. In acyclic causality models, neither the centering axiom nor the law of conditional excluded middle holds for causal (or backtracking) counterfactuals (even if we restrict ourselves to the language of the generalized version of Hitchcock & Halpern 2010). (See the appendix.) What continues to hold are theorems 12 and 13. This is so also if typicality is represented by a pre-ordering.

Theorem 15. *If acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure, then* C *is focused on actuality if, and only if, for all possible worlds* w^+*,* w*, and* w' *in* $\mathcal{W}$*: if* w *and* w' *agree on the value of every variable except for exactly one variable that is (actually) exogenous in* w^+ *and* w *strongly dominates* w' *in terms of focus on* w^+*, then* $\varrho_{w^+}(w) < \varrho_{w^+}(w')$*.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

Theorem 16. *Acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure if, and only if, the contextualization of* C *respects the causal structure and is focused on actuality.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

As stated, theorems 15 and 16 are true, but quite misleading because of the following result which holds also if typicality is represented by a pre-ordering.

Theorem 17. *Every acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure.*

PROOF: See the appendix to chapter 9. *Q.E.D.*

Among others, this means that, in a sense, our first constraint never constrained extended acyclic causal models. Acyclic causal models correspond to acyclic causality models that are structured through and through. These, like all acyclic causality models, respect the causal structure.

Causality models can be further generalized by replacing the specification of the variables $\langle \mathcal{X}, R\rangle$ by a non-empty set W of possible worlds that are taken as primitive and an interpretation function $[\![\,]\!]$ that assigns to each variable in some set of variables (atomic sentence of some formal language) its (truth-) value. (For the sake of simplicity I will make the functions e_w part of $[\![\,]\!]$ in what follows.)

$C = \langle W, (\varrho_w)_{w \in W}, [\![]\!]^e \rangle$ is a *model of causality* for the set of (propositional) variables $\mathcal{X}$ if, and only if, W is a non-empty set of possible worlds, ϱ_w is a ranking function on the power-set of W, and $[\![]\!]^e$ is an extended interpretation function with co-domain $\mathcal{R} \times \{\uparrow, \downarrow\}$ that assigns to each variable X in $\mathcal{X}$ its (truth-) value in w, as well as whether X is definitely exogenous $\uparrow$ or potentially endogenous $\downarrow$ in w. The function R can then be defined as $R(X) = \{x \in \mathcal{R} : \exists w \exists y ([\![X, w]\!]^e = \langle x, y \rangle)\}$. In models of causality, but not causality models, two possible worlds can agree on the value of every variable, but differ in the distinction between exogenous and endogenous variables or the assignment of typicality (which, as always, can be represented also by a pre-ordering).

I introduce models of causality primarily for the sake of completeness. I do so with hesitancy, though, despite the fact that theorems 15 and 16 hold also for models of causality (see the appendix). The reason is the following. Once a specification of the values of all variables does not anymore determine a unique possible world, the relata of the relation of co-determination need to be enriched: they need to comprise not only variables, but also relations (of co-determination) among variables. Otherwise, co-determination fails to capture dependencies or structure that can be expressed by iterated counterfactuals. Enriching the relata in turn requires our constraints to be reformulated, especially the concepts of causal law, violation of causal law, and ancestry. This is reflected in the fact that models of causality do not anymore respect the causal structure even if they are acyclic (in the obvious sense): theorem 17 does not hold if causality models are generalized to models of causality (again, see the appendix).

To illustrate, suppose exogenous variable L takes on value 1 if a particular material used in construction is lit, and 0 otherwise; endogenous variable F takes value 1 if this material catches fire, and 0 otherwise; exogenous variable S takes on value 1 if this material is for sale, and 0 otherwise; and endogenous variable H takes on value 1 if this material is sold for a high price, and 0 otherwise. Without presupposing the conditional theory of dispositions, suppose the material is non-flammable if, and only if, it would not catch fire just in case it were lit: $F = 1 - L$ (in the notation of structural equations); and it is highly valuable if, and only if, it would be sold for a high price just in case it were for sale: $H = S$ (where these structural equations stand for sets of counterfactuals). Finally, suppose the material would be highly valuable if, and only if, it were non-flammable. Co-determination fails to capture the latter dependency or structure between non-flammability and value if its relata are restricted to the variables L, F, S, and H and, as I assume, one does not identify the (non-) flammability and value of the material with an assignment of values to these four variables.

To capture this dependency or structure, I suggest to acknowledge the nested character of modality, as in chapter 7, and proceed as follows. Start with factual variables that generate factual worlds. Then add first-order modal variables that generate first-order modal worlds (which specify whether the factual variables are exogenous or endogenous, as well as how typical their values are by specifying how typical factual worlds are). If need be, add second-order modal variables that generate second-order modal worlds (which specify whether first-order variables are exogenous or endogenous, as well as how typical their values are by specifying how typical first-order modal worlds are). Etc. This brings us back to causality models where two possible worlds differ only if they differ in the value of at least one variable, except that we now distinguish between different levels of variables.

In our example, we add the second-order exogenous variable N that takes on value 1 if the material is non-flammable, and 0 otherwise; and the second-order endogenous variable V that takes on values 1, 1/2, and 0, respectively, if the material is highly, somewhat, and not at all valuable, respectively. By filling in the right details, N co-determines V in some first-order modal world. We can also allow for co-determination among variables at different levels. Suppose a certain piece of art is highly valuable if, and only if, it were crafted by a certain artist, and someone likes this piece of art if, and only if, it is highly valuable. The former case can be modeled as one of a factual variable co-determining a first-order modal variable, and conversely for the latter.

To conclude this section and chapter, let us show that we can define soft and hard interventions in acyclic causality models, and that we can do so without possible states and truth-makers (Briggs 2012, Fine 2012b) and the costs these generate (Embry 2014), as well as without introducing a new intervention-variable and new causality model. Besides typicality represented by a ranking function or pre-ordering over a non-empty set of possible worlds, all we need is the causal distinction between exogenous and endogenous variables.

Even more terminology. Variable Y is a child of variable X if, and only if, X is a parent of Y. Variable X is a 0-th generation descendant of itself. For any natural number $k \geq 1$, variable Y is a k-th generation descendant of variable X if, and only if, Y is a child of a $(k-1)$-th generation descendant of Y. Variable Y is a non-descendant of variable X if, and only if, for every natural number $k \geq 0$: Y is not a k-th generation descendant of X.

Consider an extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$, where the assignment of typicality matters only once we appeal to theorem 9. Relative to context $\vec{u}$, endogenous variable V is set to value v by a hard intervention in possible world w if, and only if, w assigns the following values to all variables:

(I1) if Y is a non-descendant of V, then $Y(w) = Y(w_{\vec{u}})$, where $Y(w_{\vec{u}})$ is the value that Y has in the unique legal possible world $w_{\vec{u}}$;

(I2) $V(w) = v$; and

(I($2 + k$)) if Y is a k-th generation descendant of V, $k \geq 1$, then $Y(w) = F_Y(\vec{w_Y})$, where F_Y represents the structural equation for Y and $\vec{w_Y}$ are the values of $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{Y\}$ in any possible world where the non-descendants of V, as well as all j-th generation descendants of V, $0 \leq j \leq k-1$, have their values according to (I1)-(I($k + 1$)).

This definition characterizes hard interventions in great detail. It is unnecessarily complicated, though, as the following reformulation shows. Relative to context $\vec{u}$, endogenous variable V is set to value v by a hard intervention in possible world w if, and only if, w is the unique solution to all structural equations in $\mathcal{M}_{V=v} = \langle \mathcal{S}_V, \mathcal{F}^{V=v} \rangle$ in $\vec{u}$.

In principle, we can extend this definition to exogenous variables, although for these there is no difference between *making* true, i.e., intervening, and *being* true. Let a U-reduced context be a specification of the values of all exogenous variables except U. Relative to U-reduced context $\vec{u}$, exogenous variable U is set to value u by a hard intervention in possible world w if, and only if, w assigns the following values to all variables:

(I1$^+$) if Y is a non-descendant of U, then $Y(w) = Y(w_{\vec{u}^*})$, where $\vec{u}^*$ is the context in which U has value u and all other exogenous variables have their values according to $\vec{u}$.

The other clauses remain the same. Then we get that, relative to possibly X-reduced context $\vec{u}$, variable X is set to value x by a hard intervention in possible world w if, and only if, w is the unique solution to all structural equations in $\mathcal{M}_{X=x} = \langle \mathcal{S}_X, \mathcal{F}^{X=x} \rangle$ in $\vec{u}_{X=x}$.

The reformulation makes it straightforward how to generalize the definition of a hard intervention from one variable to several variables. It also suggests still further reformulations such as the following. Relative to possibly X-reduced context $\vec{u}$, variable X is set to value x by a hard intervention in possible world w if, and only if, for all variables Y: if Y is a k-th generation descendant of X, $k \geq 1$, $ND(X)$ are the non-descendants of X, and

$$X = x \wedge \vec{ND(X)} = \vec{w_{\vec{u}^*}} \rightarrow Y = y$$

is true in $\mathcal{M}$ in $\vec{u}$, then Y has value y in w.

Since

$$X = x \text{ ⫤ } \overrightarrow{ND}(X) = \vec{w}_{\vec{u}^*}$$

is true in $\mathcal{M}$ in $\vec{u}$ (where, if X is exogenous, $\vec{u}^*$ is the context in which X has value x and all other exogenous variables have their values according to $\vec{u}$, and, if X is endogenous, $\vec{u}^* = \vec{u}$), this in turn holds if, and only if, for all variables Y: if

$$X = x \text{ ⫤ } Y = y$$

is true in $\mathcal{M}$ in $\vec{u}$, then Y has value y in w. The latter in turn holds if, and only if, for all Boolean combinations of atomic sentences β: if

$$X = x \text{ ⫤ } \beta$$

is true in $\mathcal{M}$ in $\vec{u}$, then β is true in w.

Now we just carry over this definition from extended acyclic causal models that respect the causal structure via the typicality models that reproduce them according to theorem 9 to acyclic causality models. Relative to possible world w^+ in acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in W}, (\varrho_w)_{w\in W}\rangle$, variable X is set to value x by a hard intervention in possible world w if, and only if, for all sentences β: if $X = x \boxdot\!\!\rightarrow \beta$ is true in w^+ in C, then β is true in w.

We can also capture soft interventions. Relative to possible world w^+ in acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in W}, (\varrho_w)_{w\in W}\rangle$, variable X_j is set to a value in $R_j \subseteq R\left(X_j\right)$ by a possibly soft intervention in possible world w if, and only if, for all sentences β: if $X_j \in R_j \boxdot\!\!\rightarrow \beta$ is true in w^+ in C, then β is true in w.

So far we have focused on interventions that constrain or influence which value a variable takes on, possibly completely determining a specific value of it. We can add that such an intervention would make a difference in w^+ if, and only if, the value X takes on in w differs from the value X takes on in w^+. It may be that an intervention does not only fail to change the value X takes on, but does not even target the value of X in w^+, but rather the causal law that governs X in w^+ or something different still. To characterize interventions in general, let α be an arbitrary sentence.

> **Intervention** Relative to possible world w^+ in acyclic causality model $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in W}, (\varrho_w)_{w\in W}\rangle$, α is made true by an intervention in possible world w if, and only if, for all sentences β: if $\alpha \boxdot\!\!\rightarrow \beta$ is true in w^+ in C, then β is true in w.

In acyclic causality models, making true is a special case of being true, a note truly made to end on.

9.9 Appendix: Proofs

Theorem 9. *Extended acyclic causal model* $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ *respects the causal structure if, and only if, there is a typicality model* $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ *with the same signature* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ *such that:*

T for every context $\vec{u}$ *in* $R(\mathcal{U})$*:* $\varrho_{\vec{u}} = \varrho_{w_{\vec{u}}}$*, where* $w_{\vec{u}}$ *is the unique legal world determined by* $\vec{u}$ *in* $\mathcal{M}$*; and*

SE $F_i \in \mathcal{F}$ *if, and only if,* F_i *holds in* $\mathcal{M}^*$ *if, and only if, for all* $\vec{w}_i$ *in* $\mathcal{W}_i$*:* $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \boxdot\!\!\rightarrow V_i = F_i(\vec{w}_i)$ *is true in all legal possible worlds* $w_{\vec{u}}$ *in* $\mathcal{M}^*$*.*

PROOF: Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model which respects the causal structure. We will construct a typicality model $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ with the same signature and the appropriate features. (I will remark in parentheses how to proceed if, for each context $\vec{u}$ in $R(\mathcal{U})$, typicality – or rather: atypicality – in context $\vec{u}$ is represented by a pre-ordering $\preceq_{\vec{u}}$ on $\mathcal{W}$ that is reflexive and transitive, but not necessarily connected or antisymmetric.)

For each context $\vec{u}$ in $R(\mathcal{U})$ there is exactly one possible world $w_{\vec{u}}$ in $\mathcal{W}$ that satisfies all structural equations represented by $\mathcal{F}$. Let $\mathcal{W}_0$ be the set of these legal possible worlds. For $w_{\vec{u}}$ in $\mathcal{W}_0$, let $\varrho_{w_{\vec{u}}} = \varrho_{\vec{u}}$. For illegal possible world w in $\mathcal{W} \setminus \mathcal{W}_0$, let ϱ_w copy an arbitrary $\varrho_{w_{\vec{u}}}$, for a $w_{\vec{u}}$ in $\mathcal{W}_0$. (If typicality is represented by a pre-ordering, let $\preceq_{w_{\vec{u}}} = \preceq_{\vec{u}}$ for legal possible world $w_{\vec{u}}$; for illegal possible world w, let $\preceq_w$ copy an arbitrary $\preceq_{w_{\vec{u}}}$, for a $w_{\vec{u}}$ in $\mathcal{W}_0$.) The typicality model $\mathcal{M}^*$ constructed in this way satisfies T (also if typicality is represented by a pre-ordering). Let us show next that it satisfies SE.

Suppose F_i represents the structural equation for endogenous variable V_i, $i = 1, \ldots, n$. F_i is eligible for V_i. We have to show that F_i holds in $\mathcal{M}^*$. This means we have to show for every possible world w in $\mathcal{W}$ that the following default conditionals are all true in w in $\mathcal{M}^*$: where $\vec{w}_i$ is in $\mathcal{W}_i = \times_{X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_i\}} R(X)$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \Rightarrow V_i = F_i(\vec{w}_i)\,.$$

Since the ϱ_ws for the illegal possible worlds w in $\mathcal{W} \setminus \mathcal{W}_0$ copy some $\varrho_{w_{\vec{u}}}$, for a legal possible world $w_{\vec{u}}$ in $\mathcal{W}_0$, it suffices to show that this holds for every legal possible world $w_{\vec{u}}$. (The same is true if typicality is represented by a pre-ordering.)

Each antecedent $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i$, for $\vec{w}_i$ in $\mathcal{W}_i$, is true in all and only the elements of the set of possible worlds $\{\langle \vec{w}_i, v_i \rangle : v_i \in R(V_i)\}$ (with the obvious

abuse of notation for possible worlds). There is exactly one v_i^* in $R(V_i)$, viz. the value F_i assigns to $\vec{w}_i$, such that $\langle \vec{w}_i, v_i^* \rangle$ does not violate the structural equation for V_i. For all other v_i in $R(V_i)$, the possible world $\langle \vec{w}_i, v_i \rangle$ violates the structural equation for the endogenous variable V_i. Hence, $V_i \in \mathcal{V}^*(\vec{w}_i, v_i) \setminus \mathcal{V}^*\left(\vec{w}_i, v_i^*\right)$ for all $v_i \neq v_i^*$. Furthermore, $\langle \vec{w}_i, v_i^* \rangle$ and $\langle \vec{w}_i, v_i \rangle$ agree on the values of all variables except for V_i.

Suppose $X \in \mathcal{V}^*\left(\vec{w}_i, v_i^*\right) \setminus \mathcal{V}^*(\vec{w}_i, v_i)$ for an arbitrary $v_i \neq v_i^*$. $\langle \vec{w}_i, v_i^* \rangle$ and $\langle \vec{w}_i, v_i \rangle$ agree on the value of X and $\langle \vec{w}_i, v_i \rangle$ does not violate the structural equation for X. So, there must be an exogenous or endogenous variable Y such that $Y \in An(X)$ and $\langle \vec{w}_i, v_i^* \rangle$ and $\langle \vec{w}_i, v_i \rangle$ do not agree on the value of Y. Since $\langle \vec{w}_i, v_i^* \rangle$ and $\langle \vec{w}_i, v_i \rangle$ agree on the values of all variables other than V_i, this variable must be V_i. That is, if $X \in \mathcal{V}^*\left(\vec{w}_i, v_i^*\right) \setminus \mathcal{V}^*(\vec{w}_i, v_i)$, then $V_i \in An(X)$. Since $V_i \in \mathcal{V}^*(\vec{w}_i, v_i) \setminus \mathcal{V}^*\left(\vec{w}_i, v_i^*\right)$ for all $v_i \neq v_i^*$, $\langle \vec{w}_i, v_i^* \rangle$ weakly Halpern-dominates $\langle \vec{w}_i, v_i \rangle$. Since, in acyclic causal models, $X \notin An(V_i)$ if $V_i \in An(X)$, and since $V_i \in \mathcal{V}^*(\vec{w}_i, v_i) \setminus \mathcal{V}^*\left(\vec{w}_i, v_i^*\right)$, $\langle \vec{w}_i, v_i \rangle$ does not weakly Halpern-dominate $\langle \vec{w}_i, v_i^* \rangle$.

$\mathcal{M}$ respects the causal structure. So, $\varrho_{w_{\vec{u}}}\left(\vec{w}_i, v_i^*\right) = \varrho_{\vec{u}}\left(\vec{w}_i, v_i^*\right) < \varrho_{\vec{u}}(\vec{w}_i, v_i) = \varrho_{w_{\vec{u}}}(\vec{w}_i, v_i)$ for all $v_i \neq v_i^*$. (If typicality is represented by a pre-ordering, our first constraint implies that $\langle \vec{w}_i, v_i^* \rangle \prec_{\vec{u}} \langle \vec{w}_i, v_i \rangle$ and, hence, $\langle \vec{w}_i, v_i^* \rangle \prec_{w_{\vec{u}}} \langle \vec{w}_i, v_i \rangle$ for all $v_i \neq v_i^*$, where, for any possible worlds w' and w'', $w' \prec w''$ if, and only if, $w' \preceq w''$ and $w'' \not\preceq w'$.) Since $V_i = F_i(\vec{w}_i)$ is true in $\langle \vec{w}_i, v_i^* \rangle$, it follows that the consequent of our default conditional is true in all maximally $\varrho_{w_{\vec{u}}}$-typical possible worlds in which its antecedent is true. (If typicality is represented by a pre-ordering, it follows that the consequent is true in all possible worlds a in which the antecedent is true and for which there is no possible world b in which the antecedent is true and which is such that $b \prec_{w_{\vec{u}}} a$.) Furthermore, this is so for all legal possible worlds and, hence, all possible worlds. Their truth conditions imply that our default conditionals are all true in all possible worlds. (If typicality is represented by a pre-ordering, this follows from section 9.6's truth conditions.)

The if-direction follows from the fact that, for each endogenous variable V_i, at most one eligible function holds in a typicality model $\mathcal{M}^*$. The reason is that two such functions F and F' differ only if there is a $\vec{w}_i$ such that $F(\vec{w}_i) \neq F'(\vec{w}_i)$. In this case the two default conditionals $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \Rightarrow V_i = F(\vec{w}_i)$ and $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \Rightarrow V_i = F'(\vec{w}_i)$ have inconsistent consequents, so cannot be jointly true at any possible world w. Furthermore, there always is at least one possible world in which the antecedent $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i$ is true and in which

the consequent $V_i = F(\vec{w}_i)$ must be true. This is a point Huber (2013: 729) fails to address. It follows from our truth-conditions for infinitely atypical antecedents if typicality is represented by a ranking function (and from the limit assumption if typicality is represented by a pre-ordering).

The same is true for the if-direction of the second if-and-only-if-claim in SE. For the only-if-direction of the latter, consider again, for an arbitrary antecedent $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i$, for $\vec{w}_i$ in $\mathcal{W}_i$, the unique possible world $\langle \vec{w}_i, v_i^* \rangle$, as well as the legal possible world $w_{\vec{u}^*}$ that agrees with it on the value of every exogenous variable. It follows from theorem 10 that $\langle \vec{w}_i, v_i^* \rangle$ is less typical than $w_{\vec{u}^*}$ in every legal possible world $w_{\vec{u}}$, unless the two are identical. Since $\langle \vec{w}_i, v_i^* \rangle$ is more typical, in every legal possible world, than all possible worlds $\langle \vec{w}_i, v_i \rangle$ for $v_i \neq v_i^*$, the truth conditions for counterfactuals imply that, for all $\vec{w}_i$ in $\mathcal{W}_i$,

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}_i \boxdot\!\!\rightarrow V_i = F_i(\vec{w}_i)$$

is true in all legal possible worlds in $\mathcal{M}^*$. (If typicality is represented by a pre-ordering, this follows from section 9.6's truth conditions for counterfactuals.)

Finally, let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model which does not respect the causal structure. This means that there are possible worlds w and w' that agree on the value of every variable except for exactly one endogenous variable V_i such that w strongly Halpern-dominates w', but w is not more typical than w' in some context $\vec{u}$. Consider the following default conditional, where $\vec{w}$ are the values of $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\}$ in w:

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w} \Rightarrow V_i = F_i(\vec{w})$$

The antecedent is true in both w and w'. w and w' agree on the value of every exogenous variable. So, every endogenous variable in $\mathcal{V}_0$ other than V_i whose structural equation is violated by w is also violated by w'. Here, $\mathcal{V}_0$ is the set of endogenous variables all of whose parents are exogenous variables, and $\mathcal{V}_q$ is the set of endogenous variables all of whose parents are in $\mathcal{U} \cup \mathcal{V}_0 \cup \ldots \cup \mathcal{V}_{q-1}$. Suppose every endogenous variable in $\mathcal{V}_{q-1}$ other than V_i or descendants of V_i whose structural equation is violated by w is also violated by w'. Since w and w' agree on the value of every variable in $\mathcal{V}_{q-1}$ other than V_i, every endogenous variable in $\mathcal{V}_q$ other than V_i or descendents of V_i whose structural equation is violated by w is also violated by w'. There are only finitely many "generations" $\mathcal{V}_q$ and their collection includes all endogenous variables. So, every endogenous variable other than V_i or descendants of V_i whose structural equation is violated by w is also violated by w'.

By assumption, there is at least one endogenous variable whose structural equation w' violates, whereas w does not violate the structural equation for this variable or any of its ancestors. So, this variable must be V_i.

The default conditional $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w} \Rightarrow V_i = F_i(\vec{w})$ is false in $w_{\vec{u}}$ in $\mathcal{M}^*$ if, and only if, there is at least one possible world in which $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}$ is true that is maximally $\varrho_{w_{\vec{u}}}$- ($\preceq_{\vec{u}}$-) typical and in which $V_i = F_i(\vec{w})$ is false. If typicality is represented by a ranking function and the antecedent $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}$ is infinitely atypical in $w_{\vec{u}}$, w' is such a possible world and we are done.

If w' is not such a possible world, then there is a possible world w'' in which $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}$ is true such that w'' is more typical in $w_{\vec{u}}$ than w'. $V_i = F_i(\vec{w})$ is false in w'' because w is the only possible world in which both $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\}$ and $V_i = F_i(\vec{w})$ are true. Furthermore, w is not more typical in $w_{\vec{u}}$ than w''. If w'' is such a possible world, we are done. Otherwise, there is a possible world w''' that is less typical in $w_{\vec{u}}$ than w'' in which $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \{V_i\} = \vec{w}$ is true and $V_i = F_i(\vec{w})$ false and that is not less typical in $w_{\vec{u}}$ than w. Since typicality satisfies (if typicality is represented by a pre-ordering: is assumed to satisfy) the limit assumption, this sequence of possible worlds terminates after finitely many steps at which point we are done.

This means that $F_i \in \mathcal{F}$, but that F_i does not hold in any typicality model $\mathcal{M}^*$ with the same signature as $\mathcal{M}$ that satisfies T. So, any typicality model $\mathcal{M}^*$ with the same signature as $\mathcal{M}$ that satisfies T violates SE. *Q.E.D.*

Theorem 10. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure if, and only if, for all possible worlds w and w' in $\mathcal{W}$ that agree on the value of every exogenous variable: if w strongly Halpern-dominates w', then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$ in all contexts $\vec{u}$ in $R(\mathcal{U})$.*

PROOF: Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model. The if-direction is an immediate consequence of our first constraint. (The same is true also if typicality is represented by a pre-ordering.) For the only-if-direction, suppose possible worlds w and w' agree on the value of every exogenous variable and w strongly Halpern-dominates w'. By assumption, $\mathcal{V}^*(w') \setminus \mathcal{V}^*(w)$ is not empty. Some variables in $\mathcal{V}^*(w') \setminus \mathcal{V}^*(w)$ are such that no other variable in it is causally downstream of them (because $\mathcal{M}$ is acyclic and because there are only finitely many variables). Let $\mathcal{V}^*_0$ be their collection and let $\mathcal{V}^*_j$ be the set of variables in $\left(\ldots\left((\mathcal{V}^*(w') \setminus \mathcal{V}^*(w)) \setminus \mathcal{V}^*_0\right) \setminus \ldots\right) \setminus \mathcal{V}^*_{j-1}$ such that no other variable in $\left(\ldots\left((\mathcal{V}^*(w') \setminus \mathcal{V}^*(w)) \setminus \mathcal{V}^*_0\right) \setminus \ldots\right) \setminus \mathcal{V}^*_{j-1}$ is causally downstream of them.

Take an arbitrary variable V^*_{01} in $\mathcal{V}^*_0$ and consider the possible world w_{01} that agrees with w on the value of every variable except V^*_{01} and with w' on the value of V^*_{01}. By assumption, w does not violate the structural equation for V^*_{01} or any of its ancestors. Since w_{01} agrees with w on the value of every variable except V^*_{01}, w_{01} must violate the structural equation for V^*_{01}. For the same reason w_{01} violates the structural equation for every endogenous variable in $\mathcal{V}^*(w)$ except, possibly, descendents of V^*_{01}. So, w strongly Halpern-dominates w_{01} and, hence, is more typical than it in every legal possible world.

Next take an arbitrary variable V^*_{02} in $\mathcal{V}^*_0 \setminus \left\{V^*_{01}\right\}$ and consider the possible world w_{02} that agrees with w_{01} on the value of every variable except V^*_{02} and with w' on the value of V^*_{02}. By assumption, w and, hence, w_{01} does not violate the structural equation for V^*_{02} or any of its ancestors. Furthermore, w_{02} violates the structural equation for V^*_{02} and V^*_{01} and all of $\mathcal{V}^*(w)$ except, possibly, descendents of V^*_{01} or V^*_{02}, while w_{01} violates the structural equation for V^*_{01} and all of $\mathcal{V}^*(w)$ except possibly descendents of V^*_{01} only. This is so because neither V_{02} nor V_{01} is causally upstream or downstream of the other or in $\mathcal{V}^*(w)$ and because w_{02} and w_{01} agree on the value of every variable except V_{02}. So, w_{01} strongly Halpern-dominates w_{02} and, hence, is more typical than it in every legal possible world.

Let p_0 be the number of variables in $\mathcal{V}^*_0$. Once we have reached w_{0p_0} so that there are no variables left in $\mathcal{V}^*_0$, we continue with the variables in $\mathcal{V}^*_1$ all of which are causally upstream of a variable in $\mathcal{V}^*_0$. Take an arbitrary variable V^*_{11} from $\mathcal{V}^*_1$ and consider the possible world w_{11} that agrees with w_{0p_0} on the value of every variable except V^*_{11} and with w' on the value of V^*_{11}. By assumption, w does not violate the structural equation for V^*_{11} or any of its ancestors. By construction, w_{0p_0} does not violate the structural equation for V^*_{11} or any of its ancestors either. Furthermore, w_{11} violates the structural equation for V^*_{11}, as well as the structural equations violated by w_{0p_0} except, possibly, those for descendants of V^*_{11}. So, w_{0p_0} strongly Halpern-dominates w_{11} and, hence, is more typical than it in every legal possible world.

Continuing in this way leaves us with a finite sequence of possible worlds that begins with w and ends with w' such that each possible world is less typical in every legal possible world than its immediate predecessor. The reason is that any two adjacent possible worlds in this sequence are such that they agree on the value of every variable except for exactly one endogenous variable, as well as such that the earlier occurring possible world strongly-Halpern dominates the later occurring possible world. (The same is true also if typicality is represented by a pre-ordering.) *Q.E.D.*

Theorem 11. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure and is focused on actuality if, and only if, there is a typicality model $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ with the same signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, R \rangle$ such that:*

T for every context $\vec{u}$ in $R(\mathcal{U})$: $\varrho_{\vec{u}} = \varrho_{w_{\vec{u}}}$, where $w_{\vec{u}}$ is the unique legal world determined by $\vec{u}$ in $\mathcal{M}$; and

C for every sentence ϕ in the language of the generalized version of Halpern & Hitchcock (2010) and context $\vec{u}$ in $R(\mathcal{U})$:
ϕ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, ϕ is true in $w_{\vec{u}}$ in $\mathcal{M}^*$.

PROOF: Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model which respects the causal structure and is focused on actuality. Construct $\mathcal{M}^* = \langle \mathcal{S}, (\varrho_w)_{w \in \mathcal{W}} \rangle$ as in the proof of theorem 9. $\mathcal{M}^*$ is a typicality model with the same signature that satisfies T (also if typicality is represented by a pre-ordering). Let us show next that it satisfies C.

Suppose ϕ is an atomic sentence of the form $X_i = x$ for some exogenous or endogenous variable X_i. If ϕ is true in $\mathcal{M}$ in context $\vec{u}$ this means that x is the value of X_i in the unique solution $w_{\vec{u}}$ to all structural equations represented by $\mathcal{F}$. But then $w_{\vec{u}} \in \{\langle u_1, \ldots, u_m, v_1, \ldots, v_n \rangle = \langle x_1, \ldots, x_{m+n} \rangle : x_i = x\}$. Conversely, if ϕ is not true in $\mathcal{M}$ in context $\vec{u}$ this means that x is not the value of X_i in the unique solution $w_{\vec{u}}$ to all structural equations represented by $\mathcal{F}$, in which case $w_{\vec{u}} \notin \{\langle x_1, \ldots, x_{m+n} \rangle : x_i = x\}$.

Next suppose ϕ is Boolean. Since negations, conjunctions, and disjunctions are defined in the same way in causal and typicality models, ϕ is true in $\mathcal{M}$ in context $\vec{u}$ if, and only if, ϕ is true in the legal possible world $w_{\vec{u}}$ in $\mathcal{M}^*$.

Finally, suppose ϕ is of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k \boxdot\!\!\rightarrow \psi$, for short: $\vec{X} = \vec{x} \boxdot\!\!\rightarrow \psi$, where ψ is Boolean and $X_1, \ldots, X_k$ are $k \geq 1$ distinct variables. Then ϕ is true in $\mathcal{M}$ in $\vec{u}$ if, and only if, ψ is true in the causal model $\mathcal{M}_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ in the context $\vec{u}_{\vec{X}=\vec{x}}$ which result from $\mathcal{M}$ and $\vec{u}$, respectively, by removing the structural equations for the endogenous variables among $X_1, \ldots, X_k$ and by the setting $X_i = x_i$, for all $i = 1, \ldots, k$. On the other hand, ϕ is true in $w_{\vec{u}}$ in $\mathcal{M}^*$ if, and only if, ψ is true in all $\varrho_{w_{\vec{u}}}$- ($\preceq_{w_{\vec{u}}}$-) minimal possible worlds in which $\vec{X} = \vec{x}$ is true because respect for the causal structure and focus on actuality imply that $w_{\vec{u}}$ is more typical in $w_{\vec{u}}$ than every other possible world. It suffices to consider the case where ψ is an atomic sentence of the form $Z_i = z$. In this case ψ is true in the first sense if, and only if, z is the value of Z_i in the unique solution $w^{\vec{X}=\vec{x}}_{\vec{u}_{\vec{X}=\vec{x}}} =: w^*$ to all structural equations represented by $\mathcal{F}^{\vec{X}=\vec{x}}$ in $\vec{u}_{\vec{X}=\vec{x}}$.

We need to show that w^* is the one and only $\varrho_{w_{\vec{u}}}$- ($\preceq_{w_{\vec{u}}}$-) minimal possible world in which $\vec{X} = \vec{x}$ is true. w^* is a possible world in which $\vec{X} = \vec{x}$ is true. It differs from any other possible world w' in which $\vec{X} = \vec{x}$ is true in at most the values for the variables in $\mathcal{U} \cup \mathcal{V} \setminus \{X_1, \ldots, X_k\}$. w^* agrees with $w_{\vec{u}}$ on the value of all exogenous variables in $\mathcal{U} \setminus \{X_1, \ldots, X_k\}$. Therefore, if a possible world w' in which $\vec{X} = \vec{x}$ is true differs from w^* in the value of an exogenous variable U, w' differs from $w_{\vec{u}}$ in the value of U. This means that, unless the two agree on the value of every exogenous variable, w^* strongly dominates any such possible world w' in terms of focus on $w_{\vec{u}}$. Focus on actuality implies that any such possible world w' is less typical in $\vec{u}$ – and, hence, $w_{\vec{u}}$ – than w^*.

This leaves possible worlds in which $\vec{X} = \vec{x}$ is true which differ from w^* in at most the values for the endogenous variables in $\mathcal{V} \setminus \{X_1, \ldots, X_k\}$. Let w' be such a possible world and suppose $X \in \mathcal{V}^*(w^*) \setminus \mathcal{V}^*(w')$. Since w^* satisfies the structural equations for all endogenous variables in $\mathcal{V} \setminus \{X_1, \ldots, X_k\}$, it must be that $X \in \{X_1, \ldots, X_k\}$. Since w' and w^* agree on the values of $X_1, \ldots, X_k$, and since, by assumption, w' satisfies the structural equation for X, there must be an exogenous or endogenous variable Y such that $Y \in An(X)$ and w' and w^* differ in the value for Y. The latter implies that Y is endogenous, but not among $X_1, \ldots, X_k$. So, w^* does not violate the structural equation for Y. We are done, if w' violates the structural equation for Y. Suppose it does not.

w^* and w' agree on the value of all exogenous variables, as well as $X_1, \ldots, X_k$. w^* satisfies the structural equations for all variables in $\mathcal{V} \setminus \{X_1, \ldots, X_k\}$. $Y \in \mathcal{V} \setminus \{X_1, \ldots, X_k\}$. Hence, if w' satisfies the structural equation for Y, there must be an exogenous or endogenous variable Z such that $Z \in An(Y) \subseteq An(X)$ and w' and w^* differ in the value of Z. As before, it follows that Z is endogenous, but not among $X_1, \ldots, X_k$, and that w^* satisfies the structural equation for Z. If w' violates the structural equation for Z, we are done. If not, there must be another endogenous variable $Z' \in An(Z) \subseteq An(Y) \subseteq An(X)$ with the same properties. Since there are only finitely many variables and the extended causal model is acyclic, there is an endogenous variable $Z^* \in An(X)$ such that w' violates the structural equation for Z^*, but w^* does not. So, w^* weakly Halpern-dominates w'.

Note that $\mathcal{V}^*(w') \setminus \mathcal{V}^*(w^*)$ is not empty, if w' differs from w^*. For suppose it is. Then all variables whose structural equation is violated by w' are variables whose structural equation is violated by w^*. Since w^* does not violate the structural equation for variables in $\mathcal{V} \setminus \{X_1, \ldots, X_k\}$, and since w' and w^* agree on the values of all exogenous variables, as well as $X_1, \ldots, X_k$, w' and w^* agree on the values for all variables and, hence, are identical.

Since, in extended causal models that are acyclic, $X \notin An(Z^*)$ if $Z^* \in An(X)$, since $Z^* \in \mathcal{V}^*(w') \setminus \mathcal{V}^*(w^*)$ for at least one endogenous variable Z^*, and since there are only finitely many variables, w' does not weakly Halpern-dominate w^*. Hence, w^* strongly Halpern-dominates w'. As $\mathcal{M}$ respects the causal structure, any such possible world w' is less typical in $\vec{u}$ and, hence, $w_{\vec{u}}$ than w^*.

Finally, suppose $\mathcal{M}$ does not respect the causal structure, or does, but is not focused on actuality. In the former case we proceed as in the proof of theorem 9: the falsity of the default conditional constructed there implies the falsity of the corresponding causal counterfactual which is a sentence in the language of the generalized version of Halpern & Hitchcock (2010). In the latter case, theorem 12 implies that there is a context $\vec{u}$ and possible worlds w and w' that agree on the value of every variable except for exactly one exogenous variable U_j such that w strongly dominates w' in terms of focus on $w_{\vec{u}}$, but w is not more typical in $\vec{u}$ than w'. Let $u_j(w_{\vec{u}})$ be the value U_j takes on in $w_{\vec{u}}$ and, hence, w, but not w'. The following interventionist counterfactual is true in $\mathcal{M}$ in $\vec{u}$, where $\vec{w}$ are the values of $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \left\{U_j\right\}$ in w:

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \left\{U_j\right\} = \vec{w} \,\square\!\!\rightarrow\, U_j = u_j(w_{\vec{u}})$$

However, the corresponding causal counterfactual

$$\mathcal{U} \cup \vec{\mathcal{V}} \setminus \left\{U_j\right\} = \vec{w} \,\boxdot\!\!\rightarrow\, U_j = u_j(w_{\vec{u}})$$

comes out as false in $w_{\vec{u}}$ in $\mathcal{M}^*$, where $\mathcal{M}^*$ is any of the typicality models with the same signature as $\mathcal{M}$ satisfying T that exist according to theorem 9. The reason is the same as the reason for the falsity of the default conditional in the proof of theorem 9: the consequent $U_j = u_j(w_{\vec{u}})$ is false in w', as well as every other possible world w'' in which $\mathcal{U} \cup \vec{\mathcal{V}} \setminus \left\{U_j\right\}$ is true that is more typical in $w_{\vec{u}}$ than w'. The limit assumption implies that, if w' is not a maximally $\varrho_{w_{\vec{u}}}$- ($\preceq_{w_{\vec{u}}}$-) typical possible world in which the antecedent is true, but the consequent is false (as it is if the antecedent is infinitely atypical), then some other possible world is. *Q.E.D.*

Theorem 12. *If extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure, then $\mathcal{M}$ is focused on actuality if, and only if, for all contexts $\vec{u}$ in $R(\mathcal{U})$ and possible worlds w and w' in $\mathcal{W}$ that agree on the value of every variable except for exactly one exogenous variable: if w strongly dominates w' in terms of focus on $w_{\vec{u}}$, then $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$.*

PROOF: Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model which respects the causal structure. The only-if-direction of the if-and-only-if claim is an immediate consequence of our second constraint. (The same is true also if typicality is represented by a pre-ordering.) For the if-direction, suppose possible worlds w and w' are such that, for some context $\vec{u}$ in $R(\mathcal{U})$, w strongly dominates w' in terms of focus on $w_{\vec{u}}$. We have to show that $\varrho_{\vec{u}}(w) < \varrho_{\vec{u}}(w')$ ($w \prec_{\vec{u}} w'$ if typicality is represented by a pre-ordering).

Consider the legal possible world $w_{\vec{u}^*}$ that agrees with w' on the value of every exogenous variable. $\mathcal{U}^*_{w_{\vec{u}}}(w_{\vec{u}^*}) = \mathcal{U}^*_{w_{\vec{u}}}(w')$. w strongly dominates $w_{\vec{u}^*}$ in terms of focus on $w_{\vec{u}}$. So, $\mathcal{U}^*_{w_{\vec{u}}}(w_{\vec{u}^*}) \setminus \mathcal{U}^*_{w_{\vec{u}}}(w)$ is not empty and w and $w_{\vec{u}^*}$ agree with $w_{\vec{u}}$ on the value of every exogenous variable in $\mathcal{U} \setminus \mathcal{U}^*_{w_{\vec{u}}}(w_{\vec{u}^*})$. Take an arbitrary variable U_1^* from $\mathcal{U}^*_{w_{\vec{u}}}(w_{\vec{u}^*}) \setminus \mathcal{U}^*_{w_{\vec{u}}}(w)$ and consider the possible world w_1 that agrees with w on the value of every variable except U_1^* and with $w_{\vec{u}^*}$ on the value of U_1^*. w strongly dominates w_1 in terms of focus on $w_{\vec{u}}$ and, hence, is more typical than it in context $\vec{u}$. Continuing in this way leaves us with a finite sequence of possible worlds that begins with w and ends with $w_{\vec{u}^*}$ such that each possible world in this sequence is less typical in $\vec{u}$ than its immediate predecessor. The reason is that any two adjacent possible worlds in this sequence are such that they agree on the value of every variable except for exactly one exogenous variable, as well as such that the former strongly dominates the latter in terms of focus on $w_{\vec{u}}$. If $w_{\vec{u}^*}$ and w' agree on the value of every endogenous variable, they are identical and we are done. Otherwise, $w_{\vec{u}^*}$ strongly Halpern-dominates w' and, hence, is more typical than it in every context, including $\vec{u}$. *Q.E.D.*

Theorem 13. *Extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ respects the causal structure if, and only if, the contextualization of $\mathcal{M}$ respects the causal structure and is focused on actuality.*

PROOF: Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ be an extended acyclic causal model which respects the causal structure.

Assume first that $\mathcal{W}$ is finite and that all rank-theoretic typicality functions $\varrho_{\vec{u}}$ are regular. In this case $\max\{\varrho_{\vec{u}}(w) : w \in \mathcal{W}, \vec{u} \in R(\mathcal{U})\} + 1$ exists and equals a finite natural number N. The contextualization of $\mathcal{M}$ results by replacing each typicality function $\varrho_{\vec{u}}$ with the typicality function $\varrho^*_{\vec{u}}$ which is defined as follows. Let $i_{\vec{u}}(w) = \varrho_{\vec{u}}(w) + k \times N$ if, and only if, w differs from $w_{\vec{u}}$ in the value of exactly k exogenous variables. Let $\varrho^*_{\vec{u}}(w) = -\min\{i_{\vec{u}}(w) : w \in \mathcal{W}\} + i_{\vec{u}}(w)$. The negative number is a normalization parameter and the entire definition is an application of Shenoy conditionalization (chapter 4).

Since possible worlds that agree on the values of all exogenous variables are shifted together, the contextualization of $\mathcal{M}$ still respects the causal structure. Since possible worlds that disagree with the legal possible $w_{\vec{u}}$ on the value of exactly k exogenous variables are shifted upwards (in the negative direction of higher atypicality) by $k \times N$, since a possible world w that dominates a possible world w' in terms of focus on $w_{\vec{u}}$ differs from $w_{\vec{u}}$ in the value of strictly fewer exogenous variables than w', and since $N > \varrho_{\vec{u}}(w) - \varrho_{\vec{u}}(w')$ and $N > \varrho_{\vec{u}}(w') - \varrho_{\vec{u}}(w)$, the contextualization of $\mathcal{M}$ is focused on actuality.

Now suppose typicality is represented by a pre-ordering. The new typicality ordering $\preceq^*_{\vec{u}}$ for context $\vec{u}$ is defined as follows: $w \prec^*_{\vec{u}} w'$ if w differs from $w_{\vec{u}}$ in the value of strictly fewer exogenous variables than w'; if w and w' differ from $w_{\vec{u}}$ in the value for the same number of exogenous variables, then $w \preceq^*_{\vec{u}} w'$ if, and only if, $w \preceq_{\vec{u}} w'$, where $\preceq_{\vec{u}}$ is the original typicality pre-ordering for context $\vec{u}$.

For any given context $\vec{u}$, this construction partitions the set of possible worlds $\mathcal{W}$ into $m + 1$ cells $C_0, \ldots, C_m$, where m is the number of exogenous variables. Within each cell the original typicality pre-ordering is kept, which continues to be reflexive and transitive, as well as satisfying the limit assumption, since the original typicality pre-ordering is so. (It continues to possess other properties of the original relation, including, if applicable, connectedness and antisymmetry.) Then the cells are lined up from C_0 to C_m. This construction also shows that we can drop the assumptions that there are only finitely many possible worlds and that all ranking functions are regular. We just have to allow ranking functions to take on transfinite ordinal numbers as values because the shifting parameter N needs to be sufficiently large, so will, in general, be transfinite. (This is the reason for the use of the left-sided subtraction.)

The converse is also straightforward. If the original extended acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F}, (\varrho_{\vec{u}})_{\vec{u} \in R(\mathcal{U})} \rangle$ does not respect the causal structure, then there are possible worlds w and w' that agree on the value of all exogenous variables and a context $\vec{u}$ such that $\varrho_{\vec{u}}(w) \geq \varrho_{\vec{u}}(w')$ ($w \not\prec_{\vec{u}} w'$), even though w strongly Halpern-dominates w'. By the construction of the contextualization of $\mathcal{M}$, this remains true.

There are alternative constructions with these properties. For instance, we can partition the set of possible worlds into $2 \times m$ cells by considering, for any given context $\vec{u}$, whether a possible world agrees with $w_{\vec{u}}$ in the value of exactly k exogenous variables including a fixed exogenous variable U, yielding cell C_k – as well as excluding U, yielding cell $C_{k+.5}$. As before, we leave the typicality ordering within each cell and line up the cells from C_0 to C_1 to $C_{1.5}$ to ... to C_m.

Against this background it is worth noting that there is no operation with these properties such that the resulting $\prec^*_{\vec{u}}$ orders w and w' differently than $\prec_{\vec{u}}$ only if w strongly dominates w' in terms of focus on $w_{\vec{u}}$ or w' strongly dominates w in terms of focus on $w_{\vec{u}}$. Suppose otherwise and consider three legal possible worlds $w_{1,3}$, w_2, and w_3 that disagree with the legal possible world $w_{\vec{u}}$ in the value of the following exogenous variables: U_1 and U_3, U_2, and U_3, respectively. Their values for the endogenous variables are specified in accordance with the structural equations of an appropriate extended acyclic causal model that respects the causal structure. Suppose that, initially, $w_{1,3} \preceq_{w_{\vec{u}}} w_2$ and $w_2 \preceq_{w_{\vec{u}}} w_3$. Since $w_{1,3}$ does not strongly dominate w_2 in terms of focus on $w_{\vec{u}}$ nor the other way round, and w_2 does not strongly dominate w_3 in terms of focus on $w_{\vec{u}}$ nor the other way round, $w_{1,3} \preceq^*_{w_{\vec{u}}} w_2$ and $w_2 \preceq^*_{w_{\vec{u}}} w_3$. Since w_3 strongly dominates $w_{1,3}$ in terms of focus on $w_{\vec{u}}$, we must have $w_3 \prec^*_{\vec{u}} w_{1,3}$. So, $\prec^*_{\vec{u}}$ fails to be transitive. In particular, this is so if typicality is represented by a ranking function and $w \prec^{(*)}_{\vec{u}} w'$ if, and only if, $\varrho^{(*)}_{\vec{u}}(w) < \varrho^{(*)}_{\vec{u}}(w')$. *Q.E.D.*

In the legal possible worlds of the typicality models which reproduce extended acyclic causal models that respect the causal structure according to theorem 9, the centering axiom 12. holds for causal, but not backtracking counterfactuals (even if we restrict ourselves to the language of the generalized version of Halpern & Hitchcock 2010).

Proof: Every legal possible world strongly Halpern-dominates every other possible world that agrees with it on the value of all exogenous variables. Every legal possible world strongly dominates every possible world that disagrees with it on the value of at least one exogenous variable in terms of focus on it. Hence, every legal possible world is more typical in itself than every other possible world. Consequently, the centering axiom 12. holds for causal counterfactuals. Finally, suppose that, in the drug example from section 9.5, the assignment of typicality is such that, in the legal possible world where all variables have value 0, this legal possible world is less typical in itself than the (also legal) possible world where P has value 1 and all other variables have value 0, but more typical than every other possible world where D has value 0. In this case the backtracking counterfactual $D = 0 \circ\!\!\rightarrow P = 0$ is false in this legal possible world, and so is the backtracking counterfactual $D = 0 \circ\!\!\rightarrow P = 1$. *Q.E.D.*

In the legal possible worlds of the typicality models which reproduce extended acyclic causal models that respect the causal structure according to theorem 9, the law of conditional excluded middle 0. holds for causal, but not backtracking counterfactuals, if we restrict ourselves to the language of the generalized version of Halpern & Hitchcock (2010).

Proof: The first claim follows from the proof of theorem 11. The second claim follows from the previous proof. *Q.E.D.*

In acyclic causality models, neither the centering axiom 12. nor the law of conditional excluded middle 0. holds for causal (or backtracking) counterfactuals (even if we restrict ourselves to the language of the generalized version of Halpern & Hitchcock 2010).

Proof: Consider two binary variables X and Y such that, in the possible world where both have value 0, X is definitely exogenous, Y is potentially endogenous, and the two possible worlds where X has value 0 are equally typical, as well as more typical than the possible world where $X = 1$ and $Y = 1$ which in turn is more typical than the possible world where $X = 1$ and $Y = 0$. The counterfactual $X = 1 \boxdot\!\!\rightarrow Y = 1$ is true in this possible world, while $X = 0 \boxdot\!\!\rightarrow Y = 1$ is not. Therefore, Y is (actually) endogenous in this possible world. The causal counterfactuals $X = 0 \boxdot\!\!\!\rightarrow Y = 0$ and $X = 0 \boxdot\!\!\!\rightarrow Y = 1$ are both false in this possible world. The reason is that all possible worlds in which $X = 0$ is true agree on the values of all variables that are (actually) exogenous in this possible world, i.e., X, as well as fail to violate the sole causal law of this possible world, i.e., $\{X = 1 \boxdot\!\!\rightarrow Y = 1, X = 1 \boxdot\!\!\rightarrow Y \in \{0,1\}, X = 0 \boxdot\!\!\rightarrow Y \in \{0,1\}\}$. *Q.E.D.*

Theorem 15. *If acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure, then* C *is focused on actuality if, and only if, for all possible worlds* w^+*,* w*, and* w' *in* $\mathcal{W}$*: if* w *and* w' *agree on the value of every variable except for exactly one variable that is (actually) exogenous in* w^+ *and* w *strongly dominates* w' *in terms of focus on* w^+*, then* $\varrho_{w^+}(w) < \varrho_{w^+}(w')$.

PROOF: Proceed as in the proof of theorem 12, except that $w_{\vec{u}^*}$ is w' itself (which means the proof of theorem 12 can be simplified), $w_{\vec{u}}$ is w^+, and the role played by exogenous variables is now played by variables that are (actually) exogenous in w^+. For models of causality, the construction does not deliver a finite sequence of individual possible worlds, but a finite sequence of mutually exclusive sets of possible worlds such that all possible worlds in any such set agree on the values of all variables that are (actually) exogenous in w^+. *Q.E.D.*

Theorem 16. *Acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure if, and only if, the contextualization of* C *respects the causal structure and is focused on actuality.*

PROOF: Proceed as in the proof of theorem 13, except that contextualization is defined relative to possible worlds w^+ rather than contexts $\vec{u}$ as follows, where I restrict myself to the case of typicality being represented by a pre-ordering. The new typicality ordering $\preceq^*_{w^+}$ for possible world w^+ is defined as follows: $w \prec^*_{w^+} w'$ if w differs from w^+ in the value of strictly fewer variables that are (actually) exogenous in w^+ than w'; if w and w' differ from w^+ in the value for the same number of variables that are (actually) exogenous in w^+, then $w \preceq^*_{w^+} w'$ if, and only if, $w \preceq_{w^+} w'$, where $\preceq_{w^+}$ is the original typicality pre-ordering for possible world w^+. This construction also applies to models of causality. *Q.E.D.*

Theorem 17. *Every acyclic causality model* $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ *respects the causal structure.*

PROOF: Let $C = \langle\langle \mathcal{X}, R\rangle, (e_w)_{w\in\mathcal{W}}, (\varrho_w)_{w\in\mathcal{W}}\rangle$ be an acyclic causality model, and let w^+, w, and w' be possible worlds in $\mathcal{W} = \times_{X\in\mathcal{X}} R(X)$. Suppose w and w' agree on the value of every variable except for exactly one variable that is (actually) endogenous in w^+ and w strongly dominates w' in w^+. We have to show that w is more typical than w' according to ϱ_{w^+} ($\preceq_{w^+}$, if typicality is represented by a pre-ordering). As we will see, this follows from what we will show: w', but not w, violates the causal law for the one variable that is (actually) endogenous in w^+ and on whose value w and w' disagree.

Since w strongly dominates w' in w^+, for each $X \in \mathcal{V}^*_{w^+}(w) \setminus \mathcal{V}_{w^+}(w')$ there is an $X' \in \mathcal{V}^*_{w^+}(w') \setminus \mathcal{V}^*_{w^+}(w)$ such that $X' \in An_{w^+}(X)$, but not conversely, so that there is at least one $X' \in \mathcal{V}^*_{w^+}(w') \setminus \mathcal{V}^*_{w^+}(w)$. Let X^* be such an X'.

Suppose first that w and w' disagree on value of X^*. Since w' violates the causal law for X^* in w^+, there is a subset R^* of $R(X^*)$ such that $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w'} \boxdot\!\!\rightarrow X^* \in R^*$ is true in w^+ and $X^* \in R^*$ is false in w', where $\vec{w'}$ are the values of $\mathcal{X} \setminus \vec{\{}X^*\}$ in w'. Since w and w' agree on the value of every variable except for X^*, $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w} \boxdot\!\!\rightarrow X^* \in R^*$ is true in w^+, where $\vec{w}$ are the values of $\mathcal{X} \setminus \vec{\{}X^*\}$ in w.

The antecedent $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w'}$ is true in all and only the elements of the set of possible worlds $\left\{\langle \vec{w'}, x^*\rangle : x^* \in R(X^*)\right\}$. Since $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w'} \boxdot\!\!\rightarrow X^* \in R^*$ is true in w^+, all possible worlds $\langle \vec{w'}, x^*\rangle$ that are minimal according to ϱ_{w^+} ($\preceq_{w^+}$) in this set are such that $x^* \in R^*$ – or, if w^+ is less typical according to ϱ_{w^+} ($\preceq_{w^+}$) than a possible world $\langle \vec{w'}, x^*\rangle$, all possible worlds $\langle \vec{w'}, x^*\rangle$ that are not less typical in w^+ than w^+ are such that $x^* \in R^*$.

Since w does not violate the causal law for X^* in w^+, $X^* \in R^*$ – while false in w' – is true in w. This implies that $\langle \vec{w'}, x \rangle$ is more typical in w^+ than $\langle \vec{w'}, x' \rangle$, where x is the value of X^* in w and x' is the value of X^* in w' so that $\langle \vec{w'}, x \rangle = w$ and $\langle \vec{w'}, x' \rangle = w'$. To see this, note that, since $X^* \in R^*$ is false in w', w' cannot be among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are minimal according to ϱ_{w^+} ($\preceq_{w^+}$). So, there is a possible world $\langle \vec{w'}, x^- \rangle$ that is more typical in w^+ than w'. If $x^- = x$, we are done. So, suppose $x^- \neq x$, consider the possible world $w^- = \langle \vec{w'}, x^- \rangle$, and note that w is among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are minimal according to ϱ_{w^+} ($\preceq_{w^+}$). Otherwise there is a subset R^- of $R(X^*)$ that excludes x such that $\mathcal{X} \setminus \vec{\{X^*\}} = \vec{w} \boxdot\!\!\rightarrow X^* \in R^-$ is true in w^+, where x is the value of X^* in w. In this case w violates the causal law for X^* in w^+. Contradiction.

Suppose first that w^+ is not less typical in w^+ than any possible world $\langle \vec{w'}, x^* \rangle$. w^- is not more typical in w^+ than w. If w^- is less typical in w^+ than w, then w' is less typical in w^+ than w and we are done. If w^- is not less typical in w^+ than w, we have $w \not\preceq_{w^+} w^-$ and $w^- \not\preceq_{w^+} w$. If $\preceq_{w^+}$ is connected, as is ϱ_{w^+}, w^- and w are equally typical in w^+ and w is more typical in w^+ than w'. If $\preceq_{w^+}$ is not connected, R^* is of the form $\{x^=\}$ for some $x^=$ in $R(X^*)$. If w^- is among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, then $x^- = x^= = x$ and we are done. If w^- is not among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, then so is some $w_0^- = \langle \vec{w'}, x_0^- \rangle$ that is more typical in w^+ than w^- and, hence, w'. The reason is that typicality is assumed to satisfy the limit assumption. Since w is also among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, $x_0^- = x^= = x$ and we are done.

Now suppose that w^+ is less typical in w^+ than a possible world $\langle \vec{w'}, x^* \rangle$. Since w is among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are minimal according to ϱ_{w^+} ($\preceq_{w^+}$), w is not less typical in w^+ than w^+. Since w' violates the causal law for X^* in w^+, w' is less typical in w^+ than w^+. If w is more typical in w^+ than w^+, then w is more typical in w^+ than w' and we are done. If w is not more typical in w^+ than w^+, we have $w \not\preceq_{w^+} w^+$ and $w^+ \not\preceq_{w^+} w$. If $\preceq_{w^+}$ is connected, as is ϱ_{w^+}, w^+ and w are equally typical in w^+ and w is more typical in w^+ than w'. If $\preceq_{w^+}$ is not connected, R^* is of the form $\{x^=\}$ for some $x^=$ in $R(X^*)$. Since w^+ is less typical in w^+ than a possible world $\langle \vec{w'}, x^* \rangle$, there is some $w_0^+ = \langle \vec{w'}, x_0^+ \rangle$ which is more typical in w^+ than w^+. If w_0^+ is among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, then $x_0^+ = x^= = x$ and we are done. If w_0^+ is not among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, then so is some $w_1^+ = \langle \vec{w'}, x_1^+ \rangle$ that is more typical in w^+ than w_0^+ and, hence, w^+ and, hence, w'. The reason is that typicality is assumed to satisfy the limit assumption. Since w is also among the possible worlds $\langle \vec{w'}, x^* \rangle$ that are $\preceq_{w^+}$-minimal, $x_1^+ = x^= = x$ and we are done.

This completes the first case in which w and w' disagree on the value of X^*.

Let us move on to the second case: w and w' agree on the value of X^*. Let $X^{\neq}$ be the variable that is (actually) endogenous in w^+ and on whose value w and w' disagree. Since w' violates the causal law for X^* in w^+, but w does not, $X^{\neq}$ is a parent of X^* in w^+. Otherwise w and w' agree on the value of every variable in $Pa_{w^+}(X^*)$. Since w' violates the causal law for X^* in w^+, there is a subset R^* of $R(X^*)$ such that $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w'} \; \Box\!\!\rightarrow X^* \in R^*$ is true in w^+ and $X^* \in R^*$ is false in w', where $\vec{w'}$ are the values of $\mathcal{X} \setminus \vec{\{}X^*\}$ in w'. The definition of co-determination and the truth conditions of counterfactuals imply that $\vec{Pa_{w^+}}(X^*) = \vec{w'} \; \Box\!\!\rightarrow X^* \in R^*$ is true in w^+, where $\vec{w'}$ are the values of $\vec{Pa_{w^+}}(X^*)$ in w. By assumption, w and w' agree on the value of every variable in $Pa_{w^+}(X^*)$. So, $\vec{Pa_{w^+}}(X^*) = \vec{w} \; \Box\!\!\rightarrow X^* \in R^*$ is true in w^+, where $\vec{w}$ are the values of $\vec{Pa_{w^+}}(X^*)$ in w. The truth conditions of counterfactuals imply that there is some assignment of values $\vec{np}$ to the variables in $\vec{NP_{w^+}}(X^*)$ such that $\vec{Pa_{w^+}}(X^*) = \vec{w} \wedge \vec{NP_{w^+}}(X^*) = \vec{np} \; \Box\!\!\rightarrow X^* \in R^*$ is true in w^+, where, for any variable $X \in \mathcal{X}$, $NP_{w^+}(X) = \mathcal{X} \setminus (Pa_{w^+}(X) \cup \{X\})$. The definition of co-determination implies that $\mathcal{X} \setminus \vec{\{}X^*\} = \vec{w} \; \Box\!\!\rightarrow X^* \in R^*$ is true in w^+. By assumption, w and w' agree on the value of X^* and w' violates the causal law for X^* in w^+. Hence, so does w. Contradiction.

w does not violate the causal law for X^* in w^+ or the causal law in w^+ for any ancestor of X^* in w^+. So, w does not violate the causal law for $X^{\neq}$ in w^+. If w' violates the causal law for $X^{\neq}$ in w^+, there is a subset $R^{\neq}$ of $R(X^{\neq})$ such that $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w'} \; \Box\!\!\rightarrow X^{\neq} \in R^{\neq}$ is true in w^+ and $X^{\neq} \in R^{\neq}$ is false in w', where $\vec{w'}$ are the values of $\mathcal{X} \setminus \vec{\{}X^{\neq}\}$ in w'. Since w and w' agree on the value of every variable except for $X^{\neq}$, $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w} \; \Box\!\!\rightarrow X^{\neq} \in R^{\neq}$ is true in w^+, where $\vec{w}$ are the values of $\mathcal{X} \setminus \vec{\{}X^{\neq}\}$ in w. Since w does not violate the causal law for $X^{\neq}$ in w^+, $X^{\neq} \in R^{\neq}$ is true in w and we proceed as in the first case except that the role of X^* is now played by $X^{\neq}$.

So, suppose w' does not violate the causal law for $X^{\neq}$ in w^+. The antecedent $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w}$ is true in all and only the elements of the set of possible worlds $\{\langle \vec{w}, x^{\neq} \rangle : x^{\neq} \in R(X^{\neq})\}$. Since w' does not violate the causal law for $X^{\neq}$ in w^+, w' is among the possible worlds $\langle \vec{w}, x^{\neq} \rangle$ that are minimal according to $\varrho_{w^+}(\preceq_{w^+})$. Otherwise there is a possible world $\langle \vec{w}, x^{-} \rangle$ that is more typical in w^+ than w' which implies that, for some subset R^- of $R(X^{\neq})$ that excludes x', $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w} \; \Box\!\!\rightarrow X^{\neq} \in R^-$ is true in w^+, where x' is the value of $X^{\neq}$ in w'. But then w' violates the causal law for $X^{\neq}$ in w^+. Contradiction.

C is acyclic. Since $X^{\neq}$ is a parent of X^* in w^+, X^* is not a parent of $X^{\neq}$ in w^+.

Consequently, by the truth conditions of counterfactuals and the definition of co-determination, for any subset $R^{\neq}$ of $R(X^{\neq})$, $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w} \boxdot\!\!\rightarrow X^{\neq} \in R^{\neq}$ is true in w^{+}, where $\vec{w}$ are the values of $\mathcal{X} \setminus \vec{\{}X^{\neq}\}$ in w, if, and only if, $\mathcal{X} \setminus \{\vec{X^{*}}, X^{\neq}\} = \vec{w} \boxdot\!\!\rightarrow X^{\neq} \in R^{\neq}$ is true in w^{+}, where $\vec{w}$ are the values of $\mathcal{X} \setminus \{\vec{X^{*}}, X^{\neq}\}$ in w.

The antecedent $\mathcal{X} \setminus \{\vec{X^{*}}, X^{\neq}\} = \vec{w}$ is true in all and only the elements of the set of possible worlds $\{\langle \vec{w}, x^{*}, x^{\neq}\rangle : x^{*} \in R(X^{*}), x^{\neq} \in R(X^{\neq})\}$. Since w' does not violate the causal law for $X^{\neq}$ in w^{+}, w' is among the possible worlds $\langle \vec{w}, x^{*}, x^{\neq}\rangle$ that are minimal according to $\varrho_{w^{+}}$ ($\preceq_{w^{+}}$). Otherwise there is a possible world $\langle \vec{w}, x^{*}_{0}, x^{\neq}_{0}\rangle$ that is more typical in w^{+} than w' which implies that, for some subset R^{-} of $R(X^{\neq})$ that excludes x', $\mathcal{X} \setminus \{\vec{X^{*}}, X^{\neq}\} = \vec{w} \boxdot\!\!\rightarrow X^{\neq} \in R^{-}$ is true in w^{+}, where x' is the value of $X^{\neq}$ in w'. But then also $\mathcal{X} \setminus \vec{\{}X^{\neq}\} = \vec{w} \boxdot\!\!\rightarrow X^{\neq} \in R^{-}$ is true in w^{+} and w' violates the causal law for $X^{\neq}$ in w^{+}. Contradiction.

The antecedent $\mathcal{X} \setminus \vec{\{}X^{*}\} = \vec{w}$ is true in all and only the elements of the set of possible worlds $\{\langle \vec{w}, x^{*}\rangle : x^{*} \in R(X^{*})\}$. w' is among the possible worlds $\langle \vec{w}, x^{*}, x^{\neq}\rangle$ that are minimal according to $\varrho_{w^{+}}$ ($\preceq_{w^{+}}$). So, w' also is among the possible worlds $\langle \vec{w}, x^{*}\rangle$ that are minimal according to $\varrho_{w^{+}}$ ($\preceq_{w^{+}}$). But then we have, for every subset R^{*} of $R(X^{*})$: if $\mathcal{X} \setminus \vec{\{}X^{*}\} = \vec{w} \boxdot\!\!\rightarrow X^{*} \in R^{*}$ is true in w^{+}, then $X^{*} \in R^{*}$ is true in w'. This contradicts the assumption that w' violates the causal law for X^{*} in w^{+}.

In sum, either w and w' disagree on the value of X^{*}, or else w and w' agree on the value of X^{*} and w', but not w, violates the causal law for $X^{\neq}$ in w^{+}. Either way, w', but not w, violates the causal law for the one variable that is (actually) endogenous in w^{+} and on whose value w and w' disagree. As we have seen in the first part, this implies that w' is less typical in w^{+} than w.

To see that this result does not also hold for all acyclic models of causality, consider the set containing the following five possible worlds: 00*a*, 00*b*, 01, 10, 11. In 00*a*, let 00*a* have typicality-rank 0, 01 typicality-rank 1, 11 typicality-rank 2, 10 typicality-rank 3, and 00*b* typicality-rank 4. Furthermore, consider two binary variables X and Y such that, in 00*a*, X is definitely exogenous and assigns value x to possible world xyz, while Y is potentially endogenous and assigns value y to possible world xyz. The sole causal law in 00*a* is:

$$\{X = 0 \boxdot\!\!\rightarrow Y = 0, X = 0 \boxdot\!\!\rightarrow Y \in \{0,1\}, X = 1 \boxdot\!\!\rightarrow Y = 1, X = 1 \boxdot\!\!\rightarrow Y \in \{0,1\}\}$$

Y is (actually) endogenous in 00*a* and has X as its sole parent in 00*a*. The possible world 00*b* does not violate any causal law in 00*a*, while the possible worlds 01 and 10 do. So, the latter two possible worlds are strongly dominated by 00*b* in 00*a*, even though both are more typical in 00*a* than 00*b*. *Q.E.D.*

As an afterthought, note that the restriction of subsets R_j of $R\left(X_j\right)$ in the definitions of causal law and violation of causal law to those that contain precisely one element cannot be lifted if typicality is represented by a pre-ordering that is not connected. The reason is the following. If typicality is represented by a pre-ordering that is not connected, then the possible world w' can violate the causal law for X^* in the possible world w^+ and fail to be $\preceq_{w^+}$-minimal (fail to be not less typical in w^+ than w^+) without being less typical in w^+ than the possible world w that does not violate the causal law for X^* in w^+ and is $\preceq_{w^+}$-minimal (is not less typical in w^+ than w^+). This is so if, in the first part of the proof of theorem 17, the possible world w_0^- (w_1^+) that is less typical in w^+ than w', as well as among the possible worlds $\langle \vec{w}', x^* \rangle$ that are $\preceq_{w^+}$-minimal, and w are incomparable with regard to their typicality in w^+. By forcing R^* to be a singleton $\{x^=\}$, we are forcing w_0^- (w_1^+) to be w. Personally, I take this to be a reason to adopt a connected representation of typicality. Together with the limit assumption, this allows one to identify the typicality of a set of possible worlds with the typicality of its most typical member(s).

Chapter 10

Belief and Counterfactuals

In this chapter I will first discuss the difference between intervening on the one hand and "observing" or conditioning on the other hand in probabilistic causal models, as well as explain why counterfactuals cannot be tested "empirically" on the similarity approach. Next I will describe how relative frequencies can be used to test chance hypotheses. Using this standard and familiar story from probability theory as role-model, I will then introduce the elements needed to test default conditionals and counterfactuals on the typicality approach: a concept, viz. absolute failure that relates to the modes of a sample; a normative principle, viz. the royal rule; and a theorem, viz. the Obvious Observation. Against the background of the previous chapter, this includes causal inference as a special case. I will conclude with a discussion of the conditions under which this testing is possible. I heavily rely on Huber (2015; ms).

10.1 Intervening and conditioning in probabilistic causal models

Recall the backtracking and causal counterfactual from sections 8.2 and 9.7:

BC If Ida had not slept in, she *must* not have had [*that* would have been *because* she did not have] wine the night before, and, *so*, she would not have gone for a run in the morning.

CC *Even* if Ida had not slept in, she would *still* have had wine the night before, and she would *still* have gone for a run in the morning.

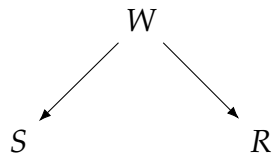

Figure 10.1: Ida doing her thing

These counterfactuals are true. They are so despite the fact that, in the actual world, Ida has wine the night before, sleeps in, and goes for a run in the morning, which makes them genuine contrary-to-fact conditionals.

The question I address in this chapter is the following: how is it possible to test or confirm these counterfactuals by information that can be obtained through observation and/or control, and to do so without assuming a particular world in a particular causality model to be true?

Before answering this question, let us first answer a different question: what differentiates a test of a backtracking counterfactual such as BC from a test of a causal counterfactual such as CC? In a nutshell, we rely on the information that it is passively observed that Ida did not sleep in when testing the backtracking counterfactual CC so that it can be assumed that this happened by being caused by its direct causes, i.e., Ida not having wine the night before. By contrast, we rely on the different information that it was actively brought about by a hard intervention that Ida did not sleep in when testing the causal counterfactual CC so that it can be assumed that one's hard intervention is the sole direct cause that this happened. More generally, when testing a causal counterfactual $\alpha \boxdot\!\!\rightarrow \beta$ we rely on the information that the antecedent α, but no logically stronger proposition, is made true by a collection of potentially soft interventions so that one's intervention is among the direct causes of the obtaining of α, possibly besides others.

On the approach of chapter 9, the difference between these tests is one of informational content. In the former case the antecedent that Ida did not sleep in is all the information we receive. In the latter case the antecedent that Ida slept in is not all the information we receive. We additionally receive the information that everything that is not causally downstream of the antecedent remains the way it was.

To get started, consider an acyclic causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$ and a probability measure $\Pr_{\mathcal{U}}$ over the power-set of $R(\mathcal{U})$. To avoid technical complications in measure theory that do not arise in the rank-theoretic setting of the later sections of this chapter, let us assume that all variables take on at most finitely many values.

We can extend $\Pr_{\mathcal{U}}$ to a unique probability measure $\Pr_{\mathcal{M}}$ over the power-set of $\mathcal{W}$ by allocating the probability of context $\vec{u}$ to the unique possible world $w_{\vec{u}}$ that is legal in $\mathcal{M}$ and assigning probability zero to all possible worlds that are illegal in $\mathcal{M}$. If the set of exogenous variables $\mathcal{U}$ is probabilistically independent in the sense of $\Pr_{\mathcal{M}}$ (section 10.5), which it is if, and only if, it is so in the sense of $\Pr_{\mathcal{U}}$, Pearl (2000/2009: 30)'s causal Markov condition theorem (section 9.6) applies: $\Pr_{\mathcal{M}}$ satisfies the causal Markov condition for the directed acyclic graph determined by $\mathcal{M}$, i.e., each variable in $\mathcal{U} \cup \mathcal{V}$ is probabilistically independent of its causal non-descendants or non-effects given its causal parents or direct causes. In this case the pair $\langle \mathcal{M}, \Pr_{\mathcal{M}} \rangle$ is Markovian; it is semi-Markovian, if the set of exogenous variables $\mathcal{U}$ is not probabilistically independent in the sense of $\Pr_{\mathcal{M}}$.

The post-intervention probability $\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}$ relative to acyclic causal model $\mathcal{M}$ after intervening on the endogenous variables $\vec{X}$ and setting their values to $\vec{x}$ can be defined to be the unique probability measure over the power-set of $\mathcal{W}$ that extends $\Pr_{\mathcal{U}}$ in a manner analogous manner to $\Pr_{\mathcal{M}}$, viz. by allocating the probability of context $\vec{u}$ to the unique possible world $w_{\vec{u}}^{\vec{X}=\vec{x}}$ that is legal in $\mathcal{M}_{\vec{X}=\vec{x}}$ and assigning probability zero to all possible worlds that are illegal in $\mathcal{M}_{\vec{X}=\vec{x}}$. It can also be calculated from the pre-intervention probability $\Pr_{\mathcal{M}}$ as follows (see Spirtes et al. 1993/2000: 51's manipulation theorem): for any possible world w in $\mathcal{W}$,

$$\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}(w) = \Pr^{*}\left(\left[\!\left[\vec{X} = X(\vec{w})\right]\!\right]\right) \times$$
$$\times \prod_{Y \in \mathcal{U} \cup \mathcal{V} \setminus \{X_1, \ldots, X_k\}} \Pr_{\mathcal{M}}\left(\left[\!\left[Y = Y(w)\right]\!\right] \mid \left[\!\left[Pa(\vec{Y}) = Pa(\vec{Y})(w)\right]\!\right]\right),$$

if the conditional probabilities in the latter product are all defined. Here, $\vec{Y}(w)$ are the values of the variables $\vec{Y}$ in w, $[\![Y = Y(w)]\!]$ is the proposition over $\mathcal{W}$ expressed by the sentence $Y = Y(w)$, and the intervention-function $\Pr^{*}$ takes on value 1 for $\vec{X}(w) = \vec{x}$ and value 0 for $\vec{X}(w) \neq \vec{x}$. (In the interest of readability, I do not distinguish between singletons of contexts or possible worlds and their elements.) The post-intervention probability satisfies the causal Markov condition for the directed acyclic graph determined by the acyclic causal model $\mathcal{M}_{\vec{X}=\vec{x}}$ if the pre-intervention probability satisfies the causal Markov condition for the directed acyclic graph that is determined by $\mathcal{M}$, i.e., if the set of exogenous variables $\mathcal{U}$ is independent in the sense of $\Pr_{\mathcal{M}}$.

Note that, for every context $\vec{u}$, as well as any two interventions on endogenous variables $\vec{X}$ and $\vec{Y}$:

$$\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}\left(w_{\vec{u}}^{\vec{X}=\vec{x}}\right) = \Pr_{\mathcal{U}}(\vec{u}) = \Pr_{\mathcal{M}_{\vec{Y}=\vec{y}}}\left(w_{\vec{u}}^{\vec{Y}=\vec{y}}\right)$$

That is, the post-intervention probability $\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}$ re-allocates the probability of context $\vec{u}$ away from the unique possible world $w_{\vec{u}}$ that is legal in $\mathcal{M}$ to the unique possible world $w_{\vec{u}}^{\vec{X}=\vec{x}}$ that is legal in $\mathcal{M}_{\vec{X}=\vec{x}}$. This means that the post-intervention probability $\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}$ is what Lewis (1976: 310) calls *the image of* the pre-intervention probability $\Pr_{\mathcal{M}}$ *on* $\vec{X} = \vec{x}$ (modulo the fact Lewis 1976 works with sentences rather than propositions). This imaging probability is the pre-intervention probability of counterfactuals $\boxright$ with antecedent $\vec{X} = \vec{x}$ which, like $\neg\exists$, validate conditional excluded middle, $\Pr_{\mathcal{M}}\left([\![\vec{X} = \vec{x} \boxright \cdot]\!]\right)$.

What I just described is a special case of Pearl (2000/2009: ch. 3)'s *do*-operator, which turns pre-intervention into post-intervention probabilities, except that it is defined also if no acyclic causal model is assumed and one is given merely a directed acyclic graph (possibly with double-arrows) and probability measure satisfying the Markov condition for it. Pearl (2017) aims at enriching the set of sentences for which the *do*-operator is defined and also notes the close relationship between intervening and imaging, though arrives at this result in a different way. In the present context of acyclic causal models, this aim amounts to enriching the set of antecedents for which interventionist counterfactuals are defined (section 9.3). Specifically, Pearl (2017) wants to allow for interventions on disjunctions (to calculate the expected utilities of disjunctive actions, among other things). This is exactly what our causality models allow for, which comprise the structure of acyclic causal models, but go beyond this structure. Pearl (2017)'s assessment that interventions on disjunctions require more structure than is present in acyclic causal models is water on the mills of the proponent of acyclic causality models. (I thank Sander Beckers for pointing me to Pearl 2017.)

Recall how we can calculate the post-intervention probability $\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}$ from the pre-intervention probability $\Pr_{\mathcal{M}}$, if, for $Y \in \mathcal{U} \cup \mathcal{V} \setminus \{X_1, \ldots, X_k\}$, the conditional probabilities

$$\Pr_{\mathcal{M}}\left([\![Y = Y(w)]\!] \mid [\![Pa(\vec{Y}) = Pa(\vec{Y})(w)]\!]\right)$$

are all defined. The latter need not be the case. $[\![Pa(\vec{Y}) = Pa(\vec{Y})(w)]\!]$ receives probability zero from $\Pr_{\mathcal{M}}$ if we intervene on $Pa(\vec{Y})$ and set them to values that they do not take on in any possible world that is legal in $\mathcal{M}$. I assume that whichever precautions are taken to side-step this issue also apply to the following considerations, as this issue does not arise in the rank-theoretic setting of the later sections. (In the present context, one can always consult the acyclic causal model, but the issue is more pressing when all one has is a directed acyclic graph and a probability measure that satisfies the causal Markov condition for it.)

These conditional probabilities take on only the extreme values 1 and 0 for endogenous variables Y; non-extreme conditional probabilities strictly between 0 and 1 are reserved for exogenous variables Y. We can rewrite the relevant equation in the following way that I have not seen elsewhere (perhaps because it holds for acyclic causal models, but, unlike the manipulation theorem, not also for pairs of directed acyclic graphs and probability measures such that the latter satisfy the causal Markov condition for the former). For any possible world w in $\mathcal{W}$,

$$\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}(w) = \Pr_{\mathcal{M}}\left([\![\vec{\mathcal{U}} = \vec{\mathcal{U}}(w)]\!]\right) \times$$
$$\times \prod_{Y \in \mathcal{V}} \Pr_{\mathcal{M}}\left([\![Y = Y(w)]\!] \mid [\![\vec{\mathcal{U}} = \vec{\mathcal{U}}(w)]\!] \cap [\![\vec{X} = \vec{x}]\!]\right).$$

In fact, this holds even if the set of exogenous variables fails to be independent in the sense of $\Pr_{\mathcal{M}}$. The conditional probabilities in the product still take on only the extreme values 1 and 0 for endogenous variables, including $X_1, \ldots, X_k$; non-extreme conditional probabilities strictly between 0 and 1 are still reserved for exogenous variables. This brings to the fore that, in acyclic causal models, the exogenous variables are causally sufficient for the endogenous variables in the sense that a specification of the former – plus the endogenous variables intervened on, if any – determines a specification of the latter.

Among others, this highlights that, in acyclic causal models, any genuinely probabilistic feature of causation among endogenous variables (that is not due to probabilistic features of the intervention) derives from probabilistic features among exogenous variables (see Papineau 2022, ms). It highlights also that, in acyclic causal models, both pre- and post-intervention probabilities satisfy the following *causal determination condition*, even if the set of exogenous variables is not independent in the sense of any of these probabilities.

> **Causal Determination Condition** Each exogenous or endogenous variable is conditionally independent of its causal non-descendants or non-effects given all of the exogenous variables, as well as all of the endogenous variables intervened on, if any.

The causal determination condition holds in acyclic causal models for the exact same reason as the causal Markov condition holds in acyclic causal models with independent exogenous variables (Pearl 2000/2009: 30, Pearl & Verma 1994: 792, Steel 2005: 22): in an acyclic causal model, the value of every variable is uniquely determined by a specification of the values of all exogenous variables plus the endogenous variables intervened on, if any.

The causal determination condition has a consequence for (causal) inference. Consider exogenous variables $U_1, \ldots, U_m$ and endogenous variables $V_1, \ldots, V_n$ and assume that they are governed by some acylic causal model or other, but it is not specified which one. Now consider what in statistics is called a marginal distribution over these variables: $\Pr := \Pr(U_1, \ldots, U_m, V_1, \ldots, V_n)$.

If we "observe" $\vec{X} = \vec{x}$ – i.e., if we receive the information that $\vec{X} = \vec{x}$ is true (and no further information) – we condition on $\vec{X} = \vec{x}$ to obtain the following new marginal distribution:

$$\Pr(U_1, \ldots, U_m, V_1, \ldots, V_n \mid x_1, \ldots, x_k)$$

By contrast, if we intervene on the endogenous variables $\vec{X}$ and set their values to $\vec{x}$ – i.e., if we receive the information that $\vec{X} = \vec{x}$ has been made true (and no further information) – we condition on $\vec{X} = \vec{x}$ *and that we are still in the same context, whichever one it is,* to obtain the following new *conditional* distribution:

$$\Pr(U_1, \ldots, U_m, V_1, \ldots, V_n \mid x_1, \ldots, x_k, U_1, \ldots, U_m)$$

This conditional distribution has the same conditions $U_1, \ldots, U_m$, no matter which acyclic causal model is true. We can use it to obtain a new marginal distribution in the following manner, where the sum ranges over all specifications $u_1, \ldots, u_m$ of the values of the exogenous variables $U_1, \ldots, U_m$, respectively:

$$\Pr^{\vec{x}} := \sum_{u_1, \ldots, u_m} \Pr(U_1, \ldots, U_m, V_1, \ldots, V_n \mid x_1, \ldots, x_k, u_1, \ldots, u_m) \Pr(u_1, \ldots, u_m)$$

If we focus on the causal Markov instead of the causal determination condition, we obtain the following conditional distribution:

$$\Pr(U_1, \ldots, U_m, V_1, \ldots, V_n \mid x_1, \ldots, x_k, Pa(X_1, \ldots, X_k))$$

The latter has different conditions $Pa(X_1, \ldots, X_k)$, even though $X_1, \ldots, X_k$ are fixed, depending on which acyclic causal model is true. To determine it, further causal assumptions are needed. If the acyclic causal model $\mathcal{M}$ is given and Pr is the pre-intervention probability $\Pr_{\mathcal{M}}$, we get the post-intervention probability $\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}}$ in this manner, where the sum now ranges over all specifications pa of the values of the causal parents or direct causes $Pa(X_1, \ldots, X_k)$ of the variables $X_1, \ldots, X_k$ intervened on:

$$\Pr_{\mathcal{M}_{\vec{X}=\vec{x}}} = \sum_{pa} \Pr(U_1, \ldots, U_m, V_1, \ldots, V_n \mid x_1, \ldots, x_k, pa) \Pr(pa)$$

Here is why. Rule 2 (action/observation exchange) of Pearl (2000/2009: sct. 3.4)'s *do*-calculus implies that, for any specification uv^{-} of the values of $UV^{-} := \{U_1, \ldots, U_m, V_1, \ldots, V_n\} \setminus (\{X_1, \ldots, X_k\} \cup Pa(X_1, \ldots, X_k))$ and any specification pa of the values of $Pa(X_1, \ldots, X_k)$, the conditional post-intervention probability of uv^{-} given pa equals its conditional pre-intervention probability given $x_1, \ldots, x_k, pa$ (because UV and $X_1, \ldots, X_k$ are d-separated by $PA(X_1, \ldots, X_k)$ after all arrows out of $X_1, \ldots, X_k$ are removed). Furthermore, both conditional probabilities of any specification of values for $\{X_1, \ldots, X_k\} \cup Pa(X_1, \ldots, X_k)$ equal 1 or both equal 0. So, for any specification uv of the values of $UV := \{U_1, \ldots, U_m, V_1, \ldots, V_n\}$ and any specification pa of the values of $Pa(X_1, \ldots, X_k)$, the conditional post-intervention probability of uv given pa equals its conditional pre-intervention probability given $x_1, \ldots, x_k, pa$. Finally, Rule 3 (insertion/deletion of actions) of Pearl (2000/2009: sct. 3.4)'s *do*-calculus implies that, for any specification pa of the values of $Pa(X_1, \ldots, X_k)$, the post-intervention probability of pa equals its pre-intervention probability (because $Pa(X_1, \ldots, X_k)$ and $X_1, \ldots, X_k$ are d-separated by the empty set after all arrows into $X_1, \ldots, X_k$ are removed). The claim then follows from the law of total probability.

The upshot of this is two-fold. First, intervening is a form of conditioning, viz. one that respects Carnap (1947b)'s "principle of total evidence" and obtains a new marginal distribution via obtaining a new conditional distribution whose condition specifies "the total evidence" received, viz. not merely that $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$ is true but the stronger claim that $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$ (is true and) has been made true. This is so even if all one has is a directed acyclic graph and probability measure satisfying the causal Markov condition for it. Second, in the context of acyclic causal models, intervening is a form of conditioning that obtains a new marginal distribution via obtaining a new conditional distribution whose condition lists all exogenous variables. Besides the classification of the variables into exogenous and endogenous, no further causal assumptions are needed.

In the following sections we will see that the situation is exactly parallel in the testing of counterfactuals, provided their semantics is given by a causality model that represents typicality by a ranking function rather than merely a pre-ordering: the difference between testing a possibly backtracking counterfactual on the one hand and a causal counterfactual on the other hand consists in the difference between conditioning to obtain a new marginal distribution directly and conditioning to obtain a new marginal distribution indirectly via obtaining a new conditional distribution whose condition lists all exogenous variables (see theorems 13 and 16). It is just that this time we consider rank-theoretic rather than probabilistic distributions.

Bibliography

[1] Andreas, Holger & Günther, Mario (2021a), A Ramsey Test Analysis of Causation for Causal Models. *British Journal for the Philosophy of Science* **72**, 587-615.

[2] Andreas, Holger & Günther, Mario (2021b), Difference-Making Causation. *Journal of Philosophy*, **118** 680-701.

[3] Andreas, Holger & Günther, Mario (2023), Actual Causation. *Dialectica*.

[4] Aristoteles (BCE/1984), *The Complete Works of Aristotle, Volumes I and II*. Ed. by J. Barnes. Princeton, NJ: Princeton University Press.

[5] Arrow, Kenneth C. (1951), *Social Choice and Individual Values*. New York: Wiley.

[6] Bar-Hillel, Maya (1980), The Base-Rate Fallacy in Probability Judgments. *Acta Psychologica* **44**, 211-233.

[7] Bar-Hillel, Yehoshua (1952), Semantic Information and Its Measures. *Transactions of the Tenth Conference on Cybernetics*. New York: Josiah Macy, Jr. Foundation, 33-48.

[8] Bar-Hillel, Yehoshua (1955), An Examination of Information Theory. *Philosophy of Science* **22**, 86-105.

[9] Bar-Hillel, Yehoshua & Carnap, Rudolf (1953), Semantic Information. *British Journal for the Philosophy of Science* **4**, 147-157.

[10] Baumgartner, Michael (2013), A Regularity Theoretic Approach to Actual Causation. *Erkenntnis* **78**, 85-109.

[11] Baumgartner, Michael & Falk, Christoph (2021), Boolean Difference-Making: A Modern Regularity Theory of Causation. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz047

[12] Bear, Adam & Knobe, Joshua (2017), Normality: Part Descriptive, Part Prescriptive. *Cognition* **167**, 25-37.

[13] Beckers, Sander (2021a), Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic* **50**, 1341-1374.

[14] Beckers, Sander (2021b), Equivalent Causal Models. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 6202-6209.

[15] Beckers, Sander & Vennekens, Joost (2017), The Transitivity and Asymmetry of Actual Causation. *Ergo* **4**, 1-27.

[16] Beckers, Sander & Vennekens, Joost (2018), A Principled Approach to Defining Actual Causation. *Synthese* **195**, 835-862.

[17] Bennett, Jonathan (2003), *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

[18] Bernstein, Allen R. & Wattenberg, Frank (1969), Non-standard Measure Theory. In W. Luxemburg (ed.), *International Symposium on the Applications of Model Theory to Algebra, Analysis, and Probability. California Institute of Technology*. New York: Holt, Reinhart, and Winston, 171-185.

[19] Briggs, Rachael (2009), The Big Bad Bug Bites Anti-Realists About Chance. *Synthese*, **167**, 81-92.

[20] Briggs, Rachael (2012), Interventionist Counterfactuals. *Philosophical Studies* **160**, 139-166.

[21] Carnap, Rudolf (1928), *Der logische Aufbau der Welt*. Berlin: Weltkreis.

[22] Carnap, Rudolf (1945), The Two Concepts of Probability. *Philosophy and Phenomenological Research* **5**, 513-532.

[23] Carnap, Rudolf (1947a), *Meaning and Necessity*. Chicago, IL: University of Chicago Press.

[24] Carnap, Rudolf (1947b), On the Application of Inductive Logic. *Philosophy and Phenomenological Research* **8**, 133-148.

[25] Carnap, Rudolf & Bar-Hillel, Yehoshua (1952), *An Outline of a Theory of Semantic Information*. Technical Report No. 247 of the Research Laboratory of Electronics, MIT.

[26] Cartwright, Nancy (1979), Causal Laws and Effective Strategies. *Noûs* **13**, 419-437.

[27] Cartwright, Nancy (1980), The Truth Doesn't Explain Much. *American Philosophical Quarterly* **17**, 159-163.

[28] Cartwright, Nancy (2007), *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. New York, NY: Cambridge University Press.

[29] Collins, John & Hall, Ned & Paul, L.A. (2004, eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

[30] Connolly, Terry & Ordóñez, Lisa D. & Coughlan, Richard (1997), Regret and Responsibility in the Evaluation of Decision Outcomes. *Organizational Behavior and Human Decision Processes* **70**, 73-85.

[31] Correa, Juan D. & Bareinboim, Elias (2020a), A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 10093-10100.

[32] Correa, Juan D. & Bareinboim, Elias (2020b), General Transportability of Soft Interventions: Completeness Results. *Proceedings of the Annual Conference on Neural Information Processing Systems* **34**.

[33] Davidson, Donald & McKinsey, J. C. C. & Suppes, Patrick (1955), Outlines of a Formal Theory of Value, I. *Philosophy of Science* **22**, 140-160.

[34] Dawid, Philip A. (1979), Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society* B **41**, 1-31.

[35] de Finetti, Bruno (1937), La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré* **7**, 1-68. Engl. Transl. by H.E. Kyburg, Jr. as "Foresight: Its Logical Laws, Its Subjective Sources."

In H.E. Kyburg, Jr. & H.E. Smokler (1964, eds.), *Studies in Subjective Probability*. New York: Wiley, 93-158.

[36] Dowe, Phil (2000), *Physical Causation*. Cambridge: Cambridge University Press.

[37] Eaton, Daniel & Murphy, Kevin (2007), Exact Bayesian Structure Learning from Uncertain Interventions.*Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, *PMLR* **2**, 107-114.

[38] Eberhardt, Frederick & Scheines, Richard (2007), Interventions and Causal Inference. *Philosophy of Science* **74**, 981-995.

[39] Edgington, Dorothy (1995), On Conditionals. *Mind* **104**, 235-329.

[40] Edgington, Dorothy (2008), Counterfactuals. *Proceedings of the Aristotelian Society* **108**, 1-21.

[41] Einstein, Albert (1905), Zur Elektrodynamik bewegter Körper. *Annalen der Physik und Chemie* **17**, 891-921.

[42] Embry, Brian (2014), Counterfactuals Without Possible Worlds? A Difficulty for Fine's Exact Semantics for Counterfactuals. *Journal of Philosophy* **111**, 276-287.

[43] Fazelpour, Sina (2021), Norms in Counterfactual Selection. *Philosophy and Phenomenological Research* **103**, 114-139.

[44] Fenton-Glynn, Luke (2017), A Proposed Probabilistic Extension of the Halpern and Pearl Definition of 'Actual Cause.' *The British Journal for the Philosophy of Science* **68**, 1061-1124.

[45] Feyerabend, Paul K. (1962), Explanation, Reduction and Empiricism. In H. Feigl & G. Maxwell (eds.), *Scientific Explanation, Space, and Time*. *Minnesota Studies in the Philosophy of Science* **3**. Minneapolis, MN: University of Minneapolis Press, 28-97.

[46] Fine, Kit (1975), Critical Notice (Review of Counterfactuals by David K. Lewis). *Mind* **84**, 451-458.

[47] Fine, Kit (2012a), A Difficulty for the Possible Worlds Analysis of Counterfactuals. *Synthese* **189**, 29-57.

[48] Fine, Kit (2012b), Counterfactuals Without Possible Worlds. *Journal of Philosophy* **109**, 221-246.

[49] Fischer, Enno (2023), Three Concepts of Actual Causation. *British Journal for the Philosophy of Science*.

[50] Foot, Philippa (1967), The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review* **5**, 1-7.

[51] Gaertner, Wulf (2009), *A Primer in Social Choice Theory*. Oxford: Oxford University Press.

[52] Galles, David & Pearl, Judea (1998). An Axiomatic Characterization of Causal Counterfactuals. *Foundations of Science* **3**, 151-182.

[53] Gallow, Dmitri J. (2021), A Model-Invariant Theory of Causation. *Philosophical Review* **130**, 45-96.

[54] Garson, James (2018), Modal Logic. In E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.

[55] Geiger, Dan & Pearl, Judea (1988), On the Logic of Causal Models. In R. Shachter, T. Levitt, L. Kanal, & J. Lemmer (eds.), *Proceedings of the Fourth Conference on Uncertainty in Artificial Intelligence*. Corvallis, OR: AUAI Press, 136-147.

[56] Gettier, Edmund L. (1963), Is Justified True Belief Knowledge? *Analysis* **23**, 121-123.

[57] Gibbard, Allan (1981), Two Recent Theories of Conditionals. In W. Harper & R. Stalnaker & G. Pearce (eds.), *Ifs*. Dordrecht: D. Reidel, 211-247.

[58] Gillies, Anthony S. (2007), Counterfactual Scorekeeping. *Linguistics and Philosophy* **30**, 329-360.

[59] Gillies, Anthony S. (2009), On Truth-Conditions for *If* (but not quite only *If*). *Philosophical Review* **118**, 325-349.

[60] Glymour, Clark (2004), Review of James Woodward, Making Things Happen: A Theory of Causal Explanation (Oxford University Press 2003). *British Journal for the Philosophy of Science* **55**, 779-790.

[61] Glymour, Clark & Danks, David & Glymour, Bruce & Eberhardt, Frederick & Ramsey, Joseph & Scheines, Richard & Spirtes, Peter & Teng, Choh Man & Zhang, Jiji (2010), Actual Causation: A Stone Soup Essay. *Synthese* **175**, 169-192.

[62] Goodman, Nelson (1954/1983), *Fact, Fiction, Forecast*. 4th ed. Cambridge, MA: Harvard University Press.

[63] Gundersen, Lars Bo (2004), Outline of a New Semantics for Counterfactuals. *Pacific Philosophical Quarterly* **85**, 1-20.

[64] Haavelmo, Trygve (1944), The Probability Approach in Econometrics. *Econometrica* **12** (Supplement), iii-vi+1-115.

[65] Hájek, Alan (ms), *Most Counterfactuals Are False*. Unpublished book manuscript.

[66] Hájek, Alan (2021), Contra Counterfactism. *Synthese* **199**, 181-210.

[67] Hall, Ned (1994), Correcting the Guide to Objective Chance. *Mind* **103**, 505-518.

[68] Hall, Ned (2000), Causation and the Price of Transitivity. *Journal of Philosophy* **97**, 198-222.

[69] Hall, Ned (2007), Structural Equations and Causation. *Philosophical Studies* **132**, 109-136.

[70] Halpern, Joseph Y. (2000). Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research* **12**, 317-337.

[71] Halpern, Joseph Y. (2003/2017), *Reasoning About Uncertainty*. 2nd ed. Cambridge, MA: MIT Press.

[72] Halpern, Joseph Y. (2008), Defaults and Normality in Causal Structures. *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, 198-208.

[73] Halpern, Joseph Y. (2013), From Causal Models to Counterfactual Structures. *The Review of Symbolic Logic* **6**, 305-322.

[74] Halpern, Joseph Y. (2016), *Actual Causality*. Cambridge, MA: MIT Press.

[75] Halpern, Joseph Y. & Hitchcock, Christopher R. (2010), Actual Causation and the Art of Modelling. In R. Dechter & H. Geffner & J. Halpern (eds.), *Heuristics, Probability, and Causality*. London: College Publications, 383-406.

[76] Halpern, Joseph Y. & Hitchcock, Christopher R. (2013), Compact Representations of Extended Causal Models. *Cognitive Science* **37**, 986-1010.

[77] Halpern, Joseph Y. & Hitchcock, Christopher R. (2015), Graded Causation and Defaults. *British Journal for the Philosophy of Science* **66**, 413-457.

[78] Halpern, Joseph Y. & Pearl, Judea (2005a), Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* **56**, 843-887.

[79] Halpern, Joseph Y. & Pearl, Judea (2005b), Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *British Journal for the Philosophy of Science* **56**, 889-911.

[80] Heckman, James J. 2005 (2005), The Scientific Model of Causality. *Sociological Methodology* **35**, 1-97.

[81] Heckman, James J. & Pinto, Rodrigo (2015), Causal Analysis After Haavelmo. *Econometric Theory* **31**, 115-151.

[82] Herzberger, Hans G. (1979), Counterfactuals and Consistency. *Journal of Philosophy* **76**, 83-88.

[83] Hiddleston, Eric (2005a), A Causal Theory of Counterfactuals. *Noûs* **39**, 232-257.

[84] Hiddleston, Eric (2005b), Causal Powers. *British Journal for the Philosophy of Science* **56**, 27-59.

[85] Hintikka, Jaakko & Pietarinen, Juhani (1966), Semantic Information and Inductive Logic. In J. Hintikka & P. Suppes (eds.), *Aspects of Inductive Logic*. Amsterdam: North-Holland, 96-112.

[86] Hitchcock, Christopher R. (2001), The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy* **98**, 273-299.

[87] Hitchcock, Christopher R. (2007), Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review* **116**, 495-532.

[88] Hitchcock, Christopher R. (2009), Structural Equations and Causation: six Counterexamples. *Philosophical Studies* **144**, 391-401.

[89] Hitchcock, Christopher R. (2018), Causal Models. In E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.

[90] Hitchcock, Christopher R. & Knobe, Joshua (2009), Cause and Norm. *The Journal of Philosophy* **106**, 587-612.

[91] Hitchcock, Christopher R. & Woodward, James (2003), Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Noûs* **37**, 181-199.

[92] Hoefer, Carl (1997), On Lewis's Objective Chance: 'Humean Supervenience Debugged'. *Mind* **106**, 321-334.

[93] Huber, Franz (2008), Assessing Theories, Bayes Style. *Synthese* **161**, 89-118.

[94] Huber, Franz (2011), Lewis Causation is a Special Case of Spohn Causation. *British Journal for the Philosophy of Science* **62**, 207-210.

[95] Huber, Franz (2012), Review of Wolfgang Spohn, The Laws of Belief: Ranking Theory and Its Philosophical Applications (Oxford University Press 2012). *Philosophy of Science* **79**, 584-588.

[96] Huber, Franz (2013), Structural Equations and Beyond. *Review of Symbolic Logic* **6**, 709-732.

[97] Huber, Franz (2014), New Foundations for Counterfactuals. *Synthese* **191**, 2167-2193.

[98] Huber, Franz (2015), What Should I Believe About What Would Have Been the Case? *Journal of Philosophical Logic* **44**, 81-110.

[99] Huber, Franz (2016), Means-End Philosophy. In W. Freitag & H. Rott & H. Sturm & A. Zinke (eds.), *Von Rang und Namen. Essays in Honour of Wolfgang Spohn*. Münster: Mentis, 173-198.

[100] Huber, Franz (2017), Why Follow The Royal Rule? *Synthese* **194**, 1565-1590.

[101] Huber, Franz (2018), *A Logical Introduction to Probability and Induction*. New York, NY: Oxford University Press.

[102] Huber, Franz (2021), *Belief and Counterfactuals. A Study in Means-End Philosophy*. New York, NY: Oxford University Press.

[103] Huber, Franz (ms), Intervening is Conditioning. Unpublished manuscript.

[104] Hume, David (1739), *A Treatise of Human Nature*. Ed. 1896 by L.A. Selby-Bigge. Oxford: Clarendon Press.

[105] Hume, David (1748), *An Enquiry Concerning Human Understanding*. Ed. 1999 by Tom L. Beauchamp. Oxford: Oxford University Press.

[106] Iatridou, Sabine (2000), The Grammatical Ingredients of Counterfactuality. *Linguistic Inquiry* **31**, 231-270.

[107] Jaber, Armin & Kocaoglu, Murat & Shanmugam, Karthikeyan & Bareinboim, Elias (2020), Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. *Advances in Neural Information Processing Systems* **33**.

[108] James, William (1896), The Will to Believe. In F. Burkhardt & F. Bowers & I. Skrupskelis (eds., 1979), *The Will to Believe and Other Essays in Popular Philosophy*. Cambridge, MA: Harvard University Press, 291-341.

[109] Jeffrey, Richard C. (1965/1983), *The Logic of Decision*. 2nd ed. Chicago, IL: University of Chicago Press.

[110] Joyce, James M. (1998), A Non-Pragmatic Vindication of Probabilism. *Philosophy of Science* **65**, 575-603.

[111] Joyce, James M. (2009), Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In F. Huber & C. Schmidt-Petri (eds.), *Degrees of Belief. Synthese Library* **342**. Dordrecht: Springer, 263-297.

[112] Kelly, Kevin T. (2007), A New Solution to the Puzzle of Simplicity. *Philosophy of Science* **74**, 561-573.

[113] Kelly, Kevin T. & Genin, Konstantin & Lin, Hanti (2016), Realism, Rhetoric, and Reliability. *Synthese* **193**, 1191-1223.

[114] Kistler, Max (2013), The Interventionist Account of Causation and Non-causal Association Laws. *Erkenntnis* **78**, 65-84.

[115] Knobe, Joshua & Nichols, Shaun (2008, eds.), *Experimental Philosophy*. Oxford: Oxford University Press.

[116] Kocaoglu, Murat & Jaber, Armin & Shanmugam, Karthikeyan & Bareinboim, Elias (2019), Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions. *Advances in Neural Information Processing Systems* **32**.

[117] Korb, Kevin B. & Hope, Lucas R. & Nicholson, Ann E. & Axnick, Karl (2004), Varieties of Causal Intervention. *Pacific Rim International Conference on AI*, 322-331.

[118] Kratzer, Angelika (1981), Partition and Revision: The Semantics of Counterfactuals. *Journal of Philosophical Logic* **10**, 201-216.

[119] Kroedel, Thomas (2020), *Mental Causation: A Counterfactual Theory*. Cambridge: Cambridge University Press.

[120] Kroedel, Thomas & Huber, Franz (2013), Counterfactual Dependence and Arrow. *Noûs* **47**, 453-466.

[121] Kuhn, Thomas S. (1962), *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

[122] Leitgeb, Hannes (2012a), A Probabilistic Semantics for Counterfactuals. Part A. *Review of Symbolic Logic* **5**, 26-84.

[123] Leitgeb, Hannes (2012b), A Probabilistic Semantics for Counterfactuals. Part B. *Review of Symbolic Logic* **5**, 85-121.

[124] Leslie, Sarah-Jane & Lerner, Adam (2016), Generic Generalizations. In E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.

[125] Lewis, David K. (1968), Counterpart Theory and Quantified Modal Logic. *Journal of Philosophy* **65**, 113-126.

[126] Lewis, David K. (1973a), *Counterfactuals*. Cambridge, MA: Harvard University Press.

[127] Lewis, David K. (1973b), Causation. *Journal of Philosophy* **70**, 556-567.

[128] Lewis, David K. (1976), Probabilities of Conditionals and Conditional Probabilities. *The Philosophical Review* **85**, 297-315.

[129] Lewis, David K. (1979), Counterfactual Dependence and Time's Arrow. *Noûs* **13**, 455-476.

[130] Lewis, David K. (1980), A Subjectivist's Guide to Objective Chance. In R.C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*. Vol. II. Berkeley: University of Berkeley Press, 263-293.

[131] Lewis, David K. (1981), Ordering Semantics and Premise Semantics for Counterfactuals. *Journal of Philosophical Logic* **10**, 217-234.

[132] Lewis, David K. (1986a), *On the Plurality of Worlds*. Oxford: Blackwell.

[133] Lewis, David K. (1986b), Introduction. In D. Lewis (1986), *Philosophical Papers II*. Oxford: Oxford University Press, ix-xvii.

[134] Lewis, David K. (1986c), Postscripts to 'Causation'. In Lewis, David K. (1986), *Philosophical Papers II*. Oxford: Oxford University Press, 172-213.

[135] Lewis, David K. (1986d), Events. In Lewis, David K. (1986), *Philosophical Papers II*. Oxford: Oxford University Press, 241-270.

[136] Lewis, David K. (1986e). Postscripts to 'Counterfactual dependence and time's arrow'. In Lewis, David K. (1986), *Philosophical Papers II*. Oxford: Oxford University Press, 52-66.

[137] Lewis, David K. (1994), Humean Supervenience Debugged. *Mind* **103**, 473-490.

[138] Lewis, David K. (2000), Causation as Influence. *Journal of Philosophy* **97**, 182-197.

[139] Mackie, John L. (1965), Causes and Conditions. *American Philosophical Quarterly* **2**, 245-264.

[140] Markowetz, Florian & Grossmann, Steffen & Spang, Rainer (2005), Probabilistic Soft Interventions in Conditional Gaussian Networks. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, *PMLR* **R5**, 214-221.

[141] Meek, Christopher & Glymour, Clark (1994), Conditioning and Intervening. *British Journal for the Philosophy of Science* **45**, 1001-1021.

[142] Menzies, Peter & Price, Huw (1993), Causation as a Secondary Quality. *British Journal for the Philosophy of Science* **44**, 187-203.

[143] Menzies, Peter (2004), Difference-Making in Context. In J. Collins & N. Hall & L.A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 139-180.

[144] Mill, John Stuart (1843), *A System of Logic, Ratiocinative and Inductive. Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. London: John W. Parker.

[145] Moore, George E. (1942), A Reply to My Critics. In P.A. Schilpp (ed.), *The Philosophy of G.E. Moore*. La Salle, IL: Open Court, 535-677.

[146] Morreau, Michael (2010), It simply does not add up: Trouble with Overall Similarity. *Journal of Philosophy* **107**, 469-490.

[147] Morris, Adam & Scott Phillips, Jonathan & Icard, Thomas & Knobe, Joshua & Gerstenberg, Tobias & Cushman, Fiery (ms), Causal Judgments Approximate the Effectiveness of Future Interventions. `https://psyarxiv.com/nq53z/`

[148] Moss, Sarah (2012), On the Pragmatics of Counterfactuals. *Noûs* **46**, 561-586.

[149] Mumford, Stephen (1998), *Dispositions*. Oxford: Oxford University Press.

[150] Newton, Isaac (1687), *Philosophiæ Naturalis Principia Mathematica*. London: Jussu Societatis Regiæ ac Typis Joseph Streater.

[151] Nozick, Robert (1981), *Philosophical Explanations*. Oxford: Oxford University Press.

[152] Okasha, Samir (2011), Theory Choice and Social Choice: Kuhn versus Arrow. *Mind* **120**, 83-115.

[153] Papineau, David (2022), The Statistical Nature of Causation. *The Monist* **105**, 247-275.

[154] Papineau, David (ms), The Causal Structure of Reality. Unpublished manuscript.

[155] Paul, L.A. (2000), Aspect Causation. *Journal of Philosophy* **97**, 235-256.

[156] Paul, L.A. & Hall, Ned (2013), *Causation: A User's Guide*. Oxford: Oxford University Press.

[157] Pearl, Judea (2000/2009), *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

[158] Pearl, Judea (2017), Physical and Metaphysical Counterfactuals: Evaluating Disjunctive Actions. *Journal of Causal Inference* **5**, 46-54.

[159] Pearl, Judea (2021), Causal and Counterfactual Inference. In Knauff, Markus & Spohn, Wolfgang (eds.), *The Handbook of Rationality*. Cambridge, MA: MIT Press, 427-438.

[160] Pearl, Judea & Verma, Thomas S. (1994), A Theory of Inferred Causation. In D. Prawitz & B. Skyrms & D. Westerståhl (eds.), *Proceedings of the Ninth International Congress of Logic, Methodology, and Philosophy of Science.* Amsterdam: Elsevier, 789-811.

[161] Pettigrew, Richard (2013), A New Epistemic Utility Argument for the Principal Principle. *Episteme* **10**, 19-35.

[162] Pollock, John (1976), The "Possible Worlds" Analysis of Counterfactuals. *Philosophical Studies* **29**, 469-476.

[163] Popper, Karl R. (1935), *Logik der Forschung. Zur Erkenntnistheorie der Modernen Naturwissenschaft*. Wien: Springer-Verlag.

[164] Price, Huw (2008), Will There Be Blood? Brandom, Hume, and the Genealogy of Modals. *Philosophical Topics* **36**, 87-97.

[165] Pust, Joel (2000), *Intuitions as Evidence*. New York, NY: Routledge.

[166] Quine, Willard Van Orman (1950), *Methods of Logic*. New York, NY: Holt, Rinehart, and Winston.

[167] Quine, Willard Van Orman (1960), *Word and Object*. Cambridge, MA: MIT Press.

[168] Quine, Willard Van Orman (1974), *The Roots of Reference*. LaSalle, IL: Open Court.

[169] Raidl, Eric (2019), Completeness for *Counter-Doxa* Conditionals – using Ranking Semantics. *Review of Symbolic Logic* **12**, 861-891.

[170] Ramsey, Frank P. (1926), Truth and Probability. In Ramsey, Frank P. (1931), *The Foundations of Mathematics and Other Logical Essays*. Ed. by R.B. Braithwaite. London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace, and Company, 156-198.

[171] Reichenbach, Hans (1956), *The Direction of Time*. Los Angeles, CA: University of California Press.

[172] Reiss, Julian (2009), Counterfactuals, Thought Experiments, and Singular Causal Analysis in History. *Philosophy of Science* **76**, 712-723.

[173] Reiter, Raymond (1980), A Logic for Default Reasoning. *Artificial Intelligence* **13**, 81-132.

[174] Roush, Sherrilyn (2005), *Tracking Truth: Knowledge, Evidence, and Science*. New York: Oxford University Press.

[175] Salmon, Wesley C. (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

[176] Salmon, Wesley C. (1984), *Causality and Explanation*. New York, NY: Oxford University Press.

[177] Sen, Amartya K. (1970), *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.

[178] Sen, Amartya K. (1986), Social Choice Theory. In K.J. Arrow & M.D. Intriligator (eds.), *Handbook of Mathematical Economics* **3**. Amsterdam: North-Holland, 1073-1181.

[179] Sider, Theodore (2002), The Ersatz Pluriverse. *Journal of Philosophy* **99**, 279-315.

[180] Skyrms, Brian (1980), *Causal Necessity. A Pragmatic Investigation of the Necessity of Laws*. New Haven, CT: Yale University Press.

[181] Skyrms, Brian (1981), Tractarian Nominalism. *Philosophical Studies* **40**, 199-206.

[182] Skyrms, Brian (1987), Dynamic Coherence and Probability Kinematics. *Philosophy of Science* **87**, 1-20.

[183] Skyrms, Brian (1988), Conditional Chance. In J.H. Fetzer (ed.), *Probability and Causality. Essays in Honor of Wesley C. Salmon*. *Synthese Library* **192**. Dordrecht: D. Reidel, 161-178.

[184] Sobel, Howard J. (1970), Utilitarianisms: Simple and General. *Inquiry* **13**, 394-449.

[185] Sober, Elliott (1975), *Simplicity*. Oxford: Clarendon Press.

[186] Sober, Elliott (1988), *Reconstructing the Past. Parsimony, Evolution, and Inference*. Cambridge, MA: MIT Press.

[187] Sober, Elliott (1990), Explanation in Biology: Let's Razor Ockham's Razor. *Royal Institute of Philosophy Supplements* **27**, 73-93.

[188] Sober, Elliott (2015), *Ockham's Razors. A User's Manual*. Cambridge: Cambridge University Press.

[189] Spirtes, Peter & Glymour, Clark & Scheines, Richard (1993/2000), *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press.

[190] Spohn, Wolfgang (1978), *Grundlagen der Entscheidungstheorie*. Kronberg/Ts.: Scriptor-Verlag.

[191] Spohn, Wolfgang (1980), Stochastic Independence, Causal Independence, and Shieldability. *Journal of Philosophical Logic* **9**, 73-99.

[192] Spohn, Wolfgang (1983a), *Eine Theorie der Kausalität*. Unpublished habilitation thesis. Munich: LMU Munich. `http://kops.uni-konstanz.de/handle/123456789/13545`

[193] Spohn Wolfgang (1983), Deterministic and Probabilistic Causes and Reasons. In C.G. Hempel & H. Putnam & W.K. Essler (eds.), *Methodology, Epistemology, and Philosophy of Science. Essays in Honour of Wolfgang Stegmüller on the Occasion of his 60th Birthday, June 3rd, 1983*. Dordrecht: D. Reidel, 371-396.

[194] Spohn, Wolfgang (1990), Direct and Indirect Causes. *Topoi* **9**, 125-145.

[195] Spohn, Wolfgang (1994), On the Properties of Conditional Independence. In P. Humphreys (ed.), *Patrick Suppes: Scientific Philosopher. Vol. 1*. Dordrecht: Kluwer, 173-196.

[196] Spohn, Wolfgang (1999), Lewis' Principal Principle ist ein Spezialfall von van Fraassens Reflexion Principle. In J. Nida-Rümelin (ed.), *Rationalität, Realismus, Revision*. Berlin: de Gruyter, 164-173.

[197] Spohn, Wolfgang (2001), Bayesian Nets Are All There Is to Causal Dependence. In M.C. Galavotti, P. Suppes & D. Costantini (eds.), *Stochastic Causality*. Stanford, CA: CSLI Publishing, 157-172.

[198] Spohn, Wolfgang (2006), Causation: An Alternative. *British Journal for the Philosophy of Science* **57**, 93-119.

[199] Spohn, Wolfgang (2010), Chance and Necessity: From Humean Supervenience to Humean Projection. In E. Eells & J. Fetzer (eds.), *The Place of Probability in Science*. *Boston Studies in the Philosophy of Science* **284**. Dordrecht: Springer, 101-131.

[200] Spohn, Wolfgang (2012), *The Laws of Belief. Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.

[201] Spohn, Wolfgang (2013), A Ranking-Theoretic Approach to Conditionals. *Cognitive Science* **37**, 1074-1106.

[202] Spohn, Wolfgang (2015), Conditionals: A Unifying Ranking-Theoretic Perspective. *Philosophers' Imprint* **15**, 1-30.

[203] Spohn, Wolfgang (2018), How the Modalities Come into the World. *Erkenntnis* **83**, 89-112.

[204] Sprenger, Jan & Hartmann, Stephan (2019), *Bayesian Philosophy of Science*. New York, NY: Oxford University Press.

[205] Stalnaker, Robert C. (1968), A Theory of Conditionals. In N. Rescher (ed.), *Studies in Logical Theory*. *American Philosophical Quarterly Monograph Series* **4**. Oxford: Blackwell, 98-112.

[206] Stalnaker, Robert C. (1981), A Defense of Conditional Excluded Middle. In W. Harper & R. Stalnaker & G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: D. Reidel, 87-104.

[207] Stalnaker, Robert C. (1994), What Is a Nonmonotonic Consequence Relation? *Fundamenta Informaticae* **21**, 7-21.

[208] Stalnaker, Robert C. (1996), Varieties of Supervenience. *Philosophical Perspectives* **10**, 221-241.

[209] Stalnaker, Robert C. (1998), On the Representation of Context. *Journal of Logic, Language, and Information* **7**, 3-19.

[210] Stalnaker, Robert C. (1999), *Context and Content*. Oxford: Oxford University Press.

[211] Steel, Daniel (2005), Indeterminism and the Causal Markov Condition. *British Journal for the Philosophy of Science* **56**, 3-26.

[212] Studený, Milan (2005), *Probabilistic Conditional Independence Structures*. London: Springer-Verlag.

[213] Suppes, Patrick (1970), *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.

[214] Swanson, Eric (2014), Ordering Supervaluationism, Counterpart Theory, and Ersatz Fundamentality. *Journal of Philosophy* **111**, 289-310.

[215] Sytsma, Justin & Livengood, Jonathan & Rose, David (2012), Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences* **43**, 814-820.

[216] Tarski, Alfred (1935), Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* **1**, 261-405.

[217] Thau, Michael (1994), Undermining and Admissibility. *Mind* **103**, 491-503.

[218] Tian, Jin & Pearl, Judea (2001), Causal Discovery from Changes. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 512-521.

[219] Tichý, Pavel (1976), A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals. *Philosophical Studies* **29**, 271-273.

[220] Tversky, Amos & Kahneman, Daniel (1982a), Evidential Impact of Base Rates. In D. Kahneman & P. Slovic & A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press, 153-160.

[221] Tversky, Amos & Kahneman, Daniel (1982b), Judgments Of and By Representativeness. In D. Kahneman & P. Slovic & A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press, 84-98.

[222] Tversky, Amos & Kahneman, Daniel (1983), Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review* **90**, 293-315.

[223] von Fintel, Kai (2001), Counterfactuals in a Dynamic Dontext. In M.J. Kenstowicz (ed.), *Ken Hale: A Life in Language*. Cambridge, MA: MIT Press, 123-152.

[224] Vranas, Peter (2004), Have Your Cake and Eat It Too: the Old Principle Reconciled with the New. *Philosophy and Phenomenological Research* **69**, 368-382.

[225] Wason, Peter C. (1966), Reasoning. In B. Foss (ed.), *New Horizons in Psychology*. Harmondsworth: Penguin, 135-151.

[226] Weber, Marcel (2021), Causal Selection versus Causal Parity in Biology: Relevant Counterfactuals and Biologically Normal Interventions. In: C.K. Waters & J. Woodward (eds.), *Philosophical Perspectives on Causal Reasoning in Biology*. *Minnesota Studies in the Philosophy of Science* **21**. Minneapolis, MN: University of Minnesota Press.

[227] Weslake, Brad (2023), A Partial Theory of Actual Causation. *British Journal for the Philosophy of Science*.

[228] Weymark, John A. (1984), Arrow's Theorem with Social Quasi-Orderings. *Public Choice* **42**, 235-246.

[229] Wilson, Alastair (2018), Metaphysical Causation. *Noûs* **52**, 723-751.

[230] Wittgenstein, Ludwig (1921), Logisch-Philosophische Abhandlung. *Annalen der Naturphilosophische* **14**, 185-262.

[231] Woodward, James (2003), *Making Things Happen. A Theory of Causal Explanation*. Oxford: Oxford University Press.

[232] Woodward, James (2018), Causation and Manipulability. In E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.

[233] Woodward, James F. (2021), *Causation With a Human Face: Normative Theory and Descriptive Psychology*. New York: Oxford University Press.

[234] Woodward, James & Hitchcock, Christopher R. (2003), Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs* **37**, 1-24.

[235] Zhang, Jiji (2013), A Lewisian Logic of Causal Counterfactuals. *Minds and Machines* **23**, 77-93.

[236] Zhang, Jiji & Lam, Wai-Yin & De Clercq, Rafael (2013), A Peculiarity in Pearl's Logic of Interventionist Counterfactuals. *Journal of Philosophical Logic* **42**, 783-794.