Statistical Mechanics and the Asymmetry of Causation

Max Heitmann

Worcester College

University of Oxford

A thesis submitted for the degree of

Bachelor of Philosophy (BPhil)

Trinity Term, 2024

Acknowledgements

I am grateful to Adam Caulton for many stimulating discussions about the ideas in this thesis, and for his insightful comments on early drafts of its core chapters. Beyond supervising me on this project, I am also fortunate to have been Adam's student throughout my undergraduate years at Oxford, and attribute a great deal of my philosophical interests, methodology, and style to his influence.

Abstract

Building on the work of David Albert in the foundations of statistical mechanics, Barry Loewer has developed a novel analysis of counterfactuals within a powerful system for naturalized metaphysics that he and Albert call "the Mentaculus" (Albert 2000; Loewer 2006). In this thesis, I apply the Mentaculus to the problem of causation. In order to relate Loewer's work on counterfactuals to an account of causation, I take my cue from the framework of structural causal models (SCMs) and the associated methods of causal-statistical inference (Pearl et al. 2016). I explain how this framework is not immediately congenial to the Mentaculus, since the former is poorly placed to handle counterfactuals with probabilistic consequents, which are central in the latter. To solve this problem, I offer a natural generalization of the formalism of SCMs to the indeterministic setting. I then use the generalized formalism to formulate a criterion for causal influence in terms of patterns of counterfactual dependence (the "intervention criterion"). Putting this criterion together with Loewer's analysis of counterfactuals, I describe a method for reconstructing a causal graph using the Mentaculus.

The payoff for this effort is twofold. First, by embedding the (time-neutral) framework of causal models within the time-asymmetric system of the Mentaculus, an explanation emerges as to why causal models can be expected to align themselves with the direction of time, with causes always preceding their effects in time. Second, by showing how a causal graph may be *reconstructed* out of the Mentaculus, a number of common assumptions about causality—usually presupposed in the causal-models formalism—can be vindicated. These include not only the presupposition of a directed and acyclic structure for the causal graph, but also the truth of a contentious principle relating causal structure to patterns of statistical independence.

Contents

[W]e may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second.

— Hume's *An Enquiry Concerning Human Understanding (7.2.29/76-77)*

1 Introduction

Causes always precede their effects in time. About this universal generalization, nearly everyone agrees. But I maintain this is a contingent fact, not a necessary truth. It is true in the actual world (and in sufficiently similar possible worlds) in virtue of various contingent features of our world, and it may well be false in other, more remote possible worlds. Since I believe the universal time-orientation of the cause-effect relation to be a contingent fact of our world, I also believe that there is a substantive enterprise concerning the explanation of this fact. In virtue of which contingent features of the actual world can we explain why causes always precede their effects? And how, exactly, does the explanation go?

These are the central questions with which this thesis is concerned. In subsequent chapters, I will attempt to trace the time orientation of causation back to its origins in the contingent structure of the world. Along the way, I'll be able to offer an explanation of various structural properties of causation (such as its directed, acyclic character, and its relation to statistics), thereby justifying the most common presuppositions of contemporary methods of causal inference. Before embarking on this project, however, a number of preliminary remarks are in order. These remarks are collected in this introductory chapter. I will first address a foundational concern regarding the motivation behind the project. Then I'll lay out some additional elements of my philosophical approach. Note that the purpose of this chapter is not so much to offer a defence of these philosophical presuppositions (beyond the scope of the present work) as it is to delineate the basic parameters of the investigation.

1.1 Why is there a Problem about the Time Direction of Causation?

First, why is there a problem about the time direction of causation at all? Is it not, is some way, *definitional* of cause and effect that the cause comes prior to the effect? Certainly, a variety of prominent views about causation, both historical and contemporary, would seem to suggest something of the kind. Most famous, perhaps, is Hume's influential "regularity theory" of causation, presented most fully in the *Enquiry*, which construes causation as merely the constant conjunction of one type of event with another, on account of which our mind forms the habit of associating the idea of one with the idea of the other in thought. But of two event-types thus observed to be constantly conjoined, and thereby associated in thought, which is to be called the cause and which the effect? Hume appears to take the answer for granted: the effect is that which follows the cause in time.

On Hume's view, then, the very distinction between cause and effect rests definitionally on a time order for the causally related events. Such views are not uncommon in contemporary analytic philosophy. A number of philosophers have attempted, in one way or another, to construe directed causation as built out of a symmetric "causal connection" relation and an asymmetric time order relation.[1] The symmetric causal connection relation might be something as simple as regular co-occurrence (as per Hume), or it might additionally incorporate certain elements from physics, for example by subsuming the regularity under a (time-symmetric) dynamical law. But however they do it, these philosophers are united in trying to account for the asymmetry of causation by grafting time asymmetry on to an underlying symmetric relation. There can therefore be no real puzzle, on these views, as to why the direction of causation is aligned with the direction of time.

In this work, I will assume all such views are misguided. In particular, I believe that what these views tend to overlook is the deep connection between causal relationships and counterfactual dependencies. This connection lies at the core of the framework known as structural causal models, perhaps the most popular formal framework for causal reasoning and inference in the contemporary social and information sciences (for a concise introduction, see Pearl et al., 2016; for a more detailed exposition, see Pearl, 2009). This is the framework that will be adopted in this thesis, and will be explained in detail in Chapter 4.

[1] See, for example, H. Reichenbach and M. Reichenbach (1999), who defines a "wide" notion of causation in this manner. The symmetric causal connection relation is provided by the time-symmetric laws of mechanics, and the time direction is supplied by 'the direction in which most thermodynamical processes in isolated systems occur' (p.127).

	No Drug	**Drug**
Low BP	81/87 recovered (93%)	234/270 recovered (87%)
High BP	192/263 recovered (73%)	55/80 recovered (69%)
Combined	273/350 recovered (78%)	289/350 recovered (83%)

Table 1.1: An instance of Simpson's paradox, taken from Pearl et al. (2016).

The connection between causation and counterfactuals endows causal relationships with an intrinsic direction—the direction of counterfactual dependence—which needn't (and, I will argue, shouldn't) be understood inherently in terms of the arrow of time. Thus, since the asymmetric aspect of causation is not explicitly sourced in the asymmetry of time, but is rather a deep intrinsic characteristic of the cause-effect relation, the regular alignment of the causal direction and the time direction begins to look considerably less trivial.

An example will help to drive home the point. Consider the dataset summarized in table 1.1. In this dataset, a correlation which holds for an entire population is reversed in each element of a chosen partition. This correlation-reversal phenomenon is known as Simpson's paradox.

In this table of data, we observe that among the whole population, a greater percentage of those taking the drug recover from their condition than those not taking the drug. However, when partitioning the population according to blood pressure (BP), the reverse effect is seen, with a greater percentage of those not taking the drug achieving recovery. So, what should we conclude: does drug taking assist recovery, or does it actually impede recovery?

What we should conclude about the efficacy of the drug depends on the causal assumptions we make about how the data were generated. In particular, a pivotal uncertainty concerns the direction of the causal relationship between drug taking and blood pressure. Suppose, on the one hand, that the direction of causal influence runs from the blood pressure to the drug taking. For instance, having high blood pressure might make a person irritable, and thus more likely either to forget to take the drug or to ignore their prescription altogether. Then the blood pressure might act as a confounding variable, potentially influencing both the probability of recovery and the probability of taking the drug. To tease out the specific causal influence of taking the drug on recovery, we would need to screen off the spurious correlation due to blood pressure. We can do so by conditionalization, which amounts to comparing the rates of recovery in each BP-sorted sub-population. Hence we conclude that the drug has a detrimental effect on recovery.

By contrast, suppose instead that the causal relationship between blood pressure and drug taking is reversed, so that it is the drug taking that has the causal effect

on blood pressure rather than vice versa. For instance, ingesting the drug might cause a lowering of the blood pressure. In this case, it would clearly be ill-advised to "screen off" the correlation due to blood pressure, since this could blind us to one of the drug's mechanisms of action. That is, perhaps the drug overall improves the odds of recovery by lowering blood pressure, though it also has a mildly toxic side effect. In this case, we can observe from the combined data that the benefits of lowered blood pressure outweigh the harms of the toxic side effect. However, if we "screen off" the drug's effect on blood pressure, we observe only the toxic side effect, as seen in the segregated data. Clearly, it is the combined data that is salient, and we conclude that the drug is beneficial.

The direction of the assumed causal relationship between blood pressure and drug taking thus leads to different inferences about the efficacy of the drug. Just one of these inferences can be correct: either the drug helps, or it hurts, but it cannot do both.[2] The correctness of one (and only one) of these inferences therefore supplies a criterion for determining the direction of the causal relationship: if the drug is efficacious in aiding recovery, then causal influence runs from drug taking to blood pressure; if the drug is detrimental to recovery, then the causal influence runs in the reverse direction, from blood pressure to drug taking.

What is the nature of this criterion? In essence, it is a counterfactual one. In the actual world described by the data in table 1.1, drug taking is positively correlated with recovery in the whole population. To say that, beyond this, drug taking causally improves the odds of recovery is to commit oneself to a variety of counterfactual claims. For instance, it is to commit oneself to the claim that the positive correlation between drug-taking and recovery would persist in a "randomized" population, where the correlation between drug taking and blood pressure is broken. Or, equivalently, it is to say that if everyone in the population had been forced to take the drug, a larger proportion of the individuals in that population would have recovered than if everyone had instead been denied access to the drug. Thus, we have in effect a counterfactual criterion for determining the direction of the "causal arrow": the arrow goes from blood pressure to drug taking just when it is the correlations seen in the segregated data would prevail in the relevant sorts of counterfactual situations, and it goes from drug taking to blood pressure if instead it is the correlations in the combined data that would prevail. For reasons that will emerge in subsequent chapters, we will call this sort of counterfactual criterion the *intervention criterion*.

[2]All things considered, that is. Of course, even if the drug is all things considered beneficial, it still has the mildly toxic side effect.

I should flag at the outset that it is far from obvious whether the intervention criterion (as briefly outlined) above can provide a "reductive" criterion for causal influence—that is, a criterion for causal influence specified in wholly non-causal terms. The issue, as has been emphasized by Woodward (1999; 2004), is that it is far from obvious whether we are in a position to specify precisely *which* counterfactuals must be considered by this criterion in the absence of tacit (or indeed explicit) reference to the surrounding causal structure. I will return to this issue in Chapter 4. For now, however, the salient point is merely that, reductive or not, the intervention criterion does not harbour a conceptual commitment to any particular time order for the events. We can perfectly well ask whether (e.g.,) the correlations seen in the combined data would have been preserved if we had (contrary to fact) forced everyone to take the drug, without making any presumptions as to the time order of blood pressure measurements and drug administrations. Nevertheless, if we learned that (e.g.,) the blood pressure measurements had always been performed *before* the drug was administered, then we'd feel confident in discarding the possibility of a causal influence of drug taking on blood pressure; consequently, we'd have to assume that it is the blood pressure that causes drug taking, and so look to the segregated data to determine the causal effect of the drug.

In sum, the direction of the causal arrow must, we presume, align with the direction of time. But given that the direction of the causal arrow can be determined by means of the intervention criterion without reference to the direction of time, it should now be clear that this presumption, however commonplace, after all cries out for a non-trivial explanation.

1.2 Naturalism, Reductionism, and the Mentaculus

In the previous section, I outlined some of the core aspects of the theory of causation that will be assumed in what follows. These aspects include a deep connection between causal relationships and counterfactual dependencies, as well as a (tentative) commitment to the formal framework of structural causal models. In this section, I lay out some of the additional philosophical presuppositions of my investigation. While I consider all of the following presuppositions plausible, I will not do much by way of arguing for them here. Rather, for present purposes they merely serve to confine the investigation to manageable proportions.

1.2.1 Naturalized Metaphysics

First, I assume naturalism. By this I mean that the "reductive base" for metaphysical analysis should consist only of the ontology and dynamics of our best theories of physics.[3] Thus, in particular, to understand the metaphysics of causation is just to explain how causal relationships can be reduced to, or constructed out of, the ingredients of an underlying physical theory.

One qualification is that I will usually pretend as though the best underlying physical theory is classical mechanics, rather than quantum mechanics. This is for simplicity's sake, as well as for better continuity with the existing literature in the foundations of statistical mechanics. I, together with several other authors (Sklar 1993, p. 12; Albert 2000, p. 16), consider this to be a largely innocuous discrepancy, insofar as the conceptual issues to be discussed translate readily from the classical to the quantum case (but see Wallace, 2023, for a dissenting view). All that notwithstanding, it would no doubt be a worthwhile extension for future work to re-trace the steps of this analysis remaining squarely within the quantum paradigm for micro-physics, to ensure that no problematic oversights have been made.

1.2.2 Time Direction is not Intrinsic

Second, I assume that time direction is not intrinsic. To put it another way, I assume that the asymmetry *of* time consists in nothing over and above the various conspicuous asymmetries of processes occurring *in* time. A helpful analogy, due originally to Boltzmann,[4] is to the apparent asymmetry of space in the vicinity of the Earth, namely the distinction between up and down. Sklar (1993, p. 389) reminds us that according to Aristotelian physics, the many manifest asymmetries of processes occurring in space—rocks always fall down, fire always rises up, we always have an immediate awareness of which direction is up etc.—are grounded in a fundamental asymmetry of space itself, which was supposed to be anisotropic with a "centre" at the centre of the Earth. Of course, nowadays we believe that all these manifest asymmetries can be explained by appeal to the local direction of the gravitational field near the surface of the Earth. Consequently, we have lost all reason to ascribe to space an intrinsic asymmetry grounding the distinction of up and down; the

[3]This statement isn't perfectly precise. For one thing, the kind of "metaphysical analysis" I have in mind concerns only the metaphysics of concrete reality; it is not intended to imply anything about (for instance) the debate between Platonists and Nominalists on the metaphysics of universals. For another, there is of course lively metaphysical debate about just what the ontologies of our best theories of physics comprise, particularly as concerns quantum mechanics.

[4]For the universe, the two directions of time are indistinguishable, just as in space there is no up or down.' (Boltzmann 1995, VII.90, originally published 1896, 1898)

distinction of up and down just consists of all the manifest asymmetries canvassed above, which are themselves already fully explained by appeal to the local direction of the gravitational field. Similarly, we might think that if we can ground all of the manifest asymmetries of processes occurring *in* time on (for example) the entropy gradient, then we have no rational warrant for maintaining that time possesses an intrinsic direction over and above this de facto entropic asymmetry.

If we accept this latter conditional, and moreover believe that the various asymmetries in time can indeed be explained in terms of some contingent facts about the structure of the world (in combination, perhaps, with the laws of nature), then the resulting picture is one of a temporal dimension in many respects like space. For if time has no intrinsic direction, then time does not, in any objective, observer-independent sense, truly *pass*. Consider: if time really passes, then it passes in some direction, and so time must have an intrinsic direction after all. Thus, in terms of the traditional philosophical distinction between the A-theory and the B-theory (McTaggart 1908), this picture of the direction of time as nothing over and above the various asymmetries of processes occurring in time is most naturally allied to the B-theory, or "block universe" view, according to which there is no metaphysically privileged present moment and events in time are most fundamentally ordered according to the *earlier than* and *later than* relations rather than by being *past*, *present* or *future*.[5]

Although it is outside the scope of the present work to give a full justification for this, it bears briefly rehearsing why such a view is plausible, especially given a prior commitment to naturalism. The crucial point is that the underlying laws of classical mechanics possess a symmetry under the reversal of time.[6] Thus, given a description of a lawful physical process, we cannot tell whether the description is in positive time or in negative time; and for any physical process allowed by the dynamical laws, the time-reverse of that process is also allowed. So the laws do not reveal an intrinsic time direction. Perhaps this is already enough, given naturalism, to dismiss the view that time direction is intrinsic.

[5]Perhaps there is also room for a view according to which, while time does not objectively pass from past to future, nevertheless time possesses an intrinsic, irreducible direction, independently of any such de facto asymmetries as the entropy gradient (on my reading, Maudlin (2002) defends this view). If such a view could be made coherent, perhaps it would deserve exclusive rights to the label of "B-theory", and the a-directional view of time I presuppose here would be better identified with McTaggart's "C-theory". For more discussion of C-theories of time, see Farr (2020).

[6]Quantum mechanics possesses a combined charge-parity-time (CPT) symmetry, rather than a pure time-reversal symmetry. I once again follow several authors in setting aside this wrinkle (Horwich 1987, p. 56; Price 1997, p. 18; Albert 2000, p. 16). In my view, the best justification for doing so is the stark implausibility that CP violations in select weak-interaction processes suffice to account for much of any of the manifest temporal asymmetries of the world.

A little more cautiously, we might argue the case as follows. Naturalism requires us to build up the description of the world out of the dynamics and ontology of our best fundamental physical theories. We've asserted that the dynamics of these theories (classical / quantum mechanics) do not possess any time asymmetry. But might not space and time be instead considered a part of the *ontology* in their own right, and thereby bring in a naturalistically acceptable time direction simply in virtue of having an intrinsically directed structure? Perhaps so; but then to suppose that our best physical theory would marry a time-reversal invariant dynamics with an intrinsically time-directed spacetime structure would be to violate the plausible general principle that our spacetime theories shouldn't postulate more structure than is required by the dynamics (this is Earman's 1992 symmetry principle SP1). Given a spacetime theory T that contains such a mismatch between spacetime symmetries and dynamical symmetries (such as Newtonian gravitation theory set in full Newtonian spacetime) we can readily imagine a theory T' that shaves off the excess spacetime structure without loss in empirical adequacy or descriptive power (in this case, Newtonian gravitation theory set in Galilean spacetime). Arguably, the latter would always be preferable to the former; so our best physical theories would never countenance such extra structure in the spacetime.

Finally, the fact that the underlying dynamical laws possess a time-reversal symmetry entails that the various asymmetries of processes in time cannot be the result of asymmetries in the underlying laws, and must instead be explained in virtue of contingent facts about the structure of the actual world—in particular, facts about the world's *boundary conditions*. It is commonly suggested, for example, that the various asymmetries of processes in time can all be traced back to the asymmetry of entropy increase, which is to be explained by appeal to a low entropy boundary condition at one temporal end of the universe, shortly after the Big Bang. While this suggestion does not, in my view, get things exactly right, I do believe that some story involving the world's boundary conditions can ultimately succeed in explaining all the process asymmetries. The details of this story will be elaborated in the next chapter. For now, I only wish to insist that the success of any story of this kind gives us good reasons to accept the antecedent of the conditional from above: we *can* ground the various process asymmetries on something other than an intrinsic asymmetry of time itself; thus, Boltzmann's analogy to the directions of space is legitimate, and we should be reluctant to ascribe to time an intrinsic direction of its own.

1.2.3 Causation as a Macroscopic Phenomenon

Third, I assume that causality, at least of the inherently asymmetric, counterfactuals-involving sort of interest to us here, is essentially a macro-level phenomenon, not present at the microscopic level of description. The basic reason for this is simple: causation has an asymmetric structure, and there are no corresponding asymmetries in the description of micro-processes onto which this structure can be mapped. It is therefore only at the macro-level that physical processes acquire the right sort of structure to be modelled causally.

This point will become particularly clear in subsequent chapters, where I will attempt to locate the asymmetry of cause and effect in the machinery of (non-equilibrium) statistical mechanics. This project obviously presupposes that the causal relationships under study relate objects or events comprising macroscopically large numbers of micro-components, for only then can the statistical-mechanical analysis even be applied. Accordingly, all the causal models in what follows are composed solely of "macro-variables" (such as the value of someone's blood pressure) rather than "micro-variables" (such as the spin on an electron).

Sourcing the origins of the causal asymmetry in statistical mechanics, including in the statistical mechanical probability measure over microstates, has the corollary that purely mechanical processes (planets orbiting, balls colliding etc.) do not exhibit causality of the appropriate sort—or rather, do not exhibit causality insofar as they are modelled purely mechanically. This calls for a few additional words of explanation.

Consider the following example, due to H. Reichenbach and M. Reichenbach (1999). A ball is thrown into the air at an angle, tracing out a parabola. Consider three points on the parabola, ABC. The equations of mechanics, being invariant under the reversal of time, do not tell us whether to describe the process in the direction ABC or CBA. It might nevertheless be supposed that this is a genuinely causal process, with an appropriately asymmetric character, and that this becomes clear when we consider the asymmetric effects of interventions. For instance, supposing I were to intervene in the flight of the ball ABC by placing a tennis racket at B. Then, we think, the ball would be diverted onto a different path ABC'. The intervention appears to have asymmetric consequences, modifying the part of the process BC' that lies to one side of the intervention but not the part AB that lies to the other side. (And, of course, we usually take for granted that the part of the process that gets modified is the part that comes *after* the intervention in time.) Thus, we might think that interventions even in purely mechanical processes have the appropriate asymmetric character to be modelled causally.

However, under closer examination, this idea is quickly found to be specious. When we say that the intervention at B leaves the path segment AB unaffected, we merely *presuppose* that interventions keep the part of the process that lies to the past of the intervention fixed when deducing the intervention's consequences. If we keep the segment AB fixed, then an intervention at B changes the subsequent path from BC to BC'. But equally, if we keep the segment BC fixed, then an intervention at B changes the preceding path from AB to A'B. Another way of putting the point is this. Let us go ahead and define "intervention" such that interventions always keep the past fixed. Then, if we choose to describe the ball's flight in negative time (as the equations of mechanics permit us to), the intervention at B must hold the reverse-time "past" CB fixed, from which it follows that the intervention results not in the path C'BA but rather CBA'. The equations of mechanics, together with our definition of intervention, tell us that the part of the ball's flight that is modified by the "intervention" depends on whether we describe the process in positive time or negative time. But since we do not know, on the basis of the mechanical equations, which is the positive time and which the negative time description, no objective asymmetry is revealed in the effects of interventions in purely mechanical processes (for more discussion on this point, see Reichenbach, 1999, pp. 43–47).

In sum, proper, asymmetric causality is an emergent phenomenon of the macroscopic domain. Its asymmetric character can only be traced to the asymmetric principles governing macroscopic processes (i.e., the "laws" of thermodynamics and kinetic theory); no analogous asymmetry arises at the microscopic level of analysis.

1.2.4 The Probability Map of the World

Fourth, and finally, I adopt a particular approach to naturalized metaphysics, which constitutes a concrete realisation of the philosophical worldview briefly canvassed above, and which has been most explicitly developed in recent work by Barry Loewer (2006; 2020; 2023) but perhaps owes its greatest intellectual debt to earlier work by David Albert (2000). This is the system that Albert & Loewer refer to as "the Mentaculus" (for an overview, see Loewer, 2020).[7]

Beyond providing a unified framework for explaining the various de facto asymmetries in time (notably: the second law of thermodynamics, the asymmetry of epistemic access to the past and future, and the asymmetry of intervention in the past and future), the Mentaculus in principle provides a full "probability map

[7]The name comes from the Coen brothers' film *A Serious Man*, in which the main character Larry's troubled brother Arthur purports to be developing a "probability map of the universe" with the same name.

of the universe": a numerical assignment of probability to any proposition about the macroscopic state of affairs at some time (henceforth, a "macro-proposition"), as well joint probabilities for all combinations of macro-propositions and their associated conditional probabilities. The next chapter recounts the historical origins of the Mentaculus, from insights emerging out of the foundations of statistical mechanics through apparently unrelated work on counterfactuals by David Lewis. It concludes with an explicit statement of the core principles of the Mentaculus and a definition of its all-encompassing "probability map".

1.3 Goals and Overview

What do I hope to achieve with this investigation? Three things. First, I hope to offer a compelling answer the basic question from section 1.1, concerning the agreement of the intervention criterion with the time-order criterion for determining the direction of causal relationships. That is, I hope to provide a satisfying explanation for the invariable alignment of the direction of causation with the direction of time. Second, by combining an adherence to the causal-models formalism as the formal means for representing causal assumptions with a commitment to physics as the ultimate reductive base for metaphysical analysis, I hope to offer a naturalistic justification for the various modelling assumptions built into the former (such as the use of DAGs, and the rule of product decomposition), thereby vindicating (or modifying, as appropriate) the formal causal-models framework. Third, and finally, by making an extended enquiry that simply *adopts* the system of the Mentaculus, I hope to shed light on its power and flexibility as a comprehensive framework for naturalized metaphysics—as well as bring out a few of its shortcomings. Thus, beyond answering the object-level questions about the explanation for the time orientation of causation and the justification for common assumptions about its structure, this investigation can also be read as an extended "stress-test" of the Mentaculus itself.

I shall now provide a brief overview of the line of argument running through the four chapters that comprise the main body of this work.

In chapter 2, I gradually build up to a full statement of the Mentaculus, arguing that it is the natural end-point of certain popular strands of thought in the foundations of statistical mechanics, and moreover that it constitutes a natural development of earlier ideas about counterfactuals and time asymmetry due to David Lewis, representing (to some extent) the fruition of his views within a naturalized metaphysics.

In chapter 3, I get into the details about the deliverances of the Mentaculus for the philosophy of time asymmetry. I argue that the Mentaculus has the resources to provide (i) an explanation for the existence of predictively successful macrodynamical equations of motion (i.e., a "macrodynamics") that work forwards in time but not backwards, and (ii) a semantics for counterfactuals, which exhibits a generic, de facto time-asymmetry that needn't be put in by hand.

These deliverances of the Mentaculus seem, tantalizingly, to correspond roughly to the two major components of formal causal modelling frameworks: the purely statistical and the interventional. After a brief survey of these elements of causal modelling and the associated methods of causal inference, Chapter 4 pursues this intuitive idea, suggesting a way to translate between statements in the language of causal models and statements in the language of the Mentaculus. It argues that the Mentaculus, via this translation scheme, provides counterfactuals-based tests on possible causal models, and contains the all the materials necessary for a procedure for constructing causal models.

Finally, Chapter 5 delivers on all this work, explaining the coincidence of the intervention criterion with the time-direction criterion for determining the direction of causal influence, as well as deriving the directed acyclic graph (DAG) structure for generic networks of causal relationships. Additionally, it offers a novel argument for the so-called "Causal Markov Condition", which is provably equivalent to the statistical rule of product decomposition in acyclic causal models, and lies at the heart of the purely statistical component of causal inference.

For the universe, the two directions of time are indistinguishable, just as in space there is no up or down.

— Boltzmann's *Lectures on Gas Theory (VII.90)*

2 Literature Review

This project works with a comprehensive framework for naturalized metaphysics known as the Mentaculus. This chapter explicates some of the core ideas, principles, and motivations behind this framework. In comes in two parts. In the first part, I give some background on (equilibrium and non-equilibrium) statistical mechanics, which provides critical context for understanding the basic shape of the Mentaculus. Then, in the second part, I explain how the Mentaculus can be viewed in some ways as a successor to David Lewis's metaphysical programme on causation, counterfactuals, and time's arrow, which however departs from Lewis in various crucial places and incorporates a number of salient lessons from the literature on the foundations of statistical mechanics.

2.1 Introduction to Statistical Mechanics

2.1.1 Basic Ideas

Statistical mechanics is concerned with the behaviour of systems comprising a very large number of components. A paradigmatic example of such a system is a classical gas in a box, which may comprise on the order of 10^{23} molecules. Each component of the system can at any time be in any one of a (finite or infinite) number of states, the set of which make up its individual state space. For example, each molecule in the gas has a six-dimensional state space of three position coordinates and three momentum coordinates. The system as a whole also has a state space, called its *phase space*, which in the classical case is simply the direct product of the individual component state-spaces (and in the quantum case, the tensor product). Thus, for

a classical gas of N molecules, the phase space is $6N$-dimensional: every point in the phase space corresponds to an assignment of position and momentum to every molecule in the gas. The Hamiltonian of the system determines the phase space trajectories via the Hamiltonian equations of motion for the generalized coordinates; thus, for any initial point x in the phase space, the Hamiltonian determines a map from that point to its t-second evolution, $\phi_t(x)$. We can also use this map to time-evolve an initial *distribution* over the phase space, giving us (in the classical case) the Liouville dynamics for the system.

So far, this is just the Hamiltonian perspective on classical mechanics.[1] Statistical mechanics begins with the introduction of a *Boltzmann coarse-graining*, a partition of the phase space into disjoint regions. (My exposition in this section loosely follows Albert (2000).) A very typical procedure is as follows. First, partition up the state space for the single component into small cells, indexed i. Since the single-component state space is the same for all the components, we can represent the state of the whole system at any time by means of N points in this single-component state space; this is called the *mu-space* representation. In this representation, each cell will be occupied, at any given time, by an integer n_i of points (representing, in the case of a gas, the number of molecules with positions and momenta within the intervals defined by the cell i). The set of these occupation numbers $\{n_1, n_2, ..., n_m\}$ is called a *distribution*. Any given distribution corresponds to a great many possible micro-configurations of the gas; indeed, in the classical case, a distribution always corresponds to uncountably many microstates, composing a continuous *region* of the phase space. The set of all distributions (with respect to a suitable partition of the mu-space) therefore furnishes a Boltzmann coarse-graining.

There are some basic constraints on how we may define the cells in our partition of mu-space. The first, and most important constraint, is that the cells in the partition must all be the *same* size (with respect to an appropriate measure over the single-component state space). Second, the size of the cells must also be tuned in a particular way: the cells should be chosen small enough such that they are effectively "point-like" to the coarse senses and measuring instruments of the human being, yet still big enough such that the number m of cells is much smaller than the number N of components (i.e., such that any given occupied cell has, on average, a *significant* occupation, $n_i \gg 1$). When the partition of mu-space satisfies these constraints, the induced Boltzmann coarse-graining constitutes a partition of the phase space into *macrostates*.

[1] See any textbook on classical mechanics, for example Goldstein et al. (2002).

To give some intuition about macrostates, consider the result of this procedure in the case of the gas. Specifying the set of occupation numbers corresponding to some small interval of position $[\mathbf{x}, \mathbf{x}+\delta\mathbf{x}]$ for all values of momentum $\mathbf{p}$ tells us something about the *density* of the gas around the position $\mathbf{x}$ (via a simple summation of the occupation numbers over all momenta), as well as about its *pressure* around position $\mathbf{x}$ (via an occupation-number weighted vector sum of the momenta), and about its *temperature* around $\mathbf{x}$ (via an occupation-number weighted average of $|\mathbf{p}|^2/2m$). Specifying these occupation numbers for *all* small intervals of position (which, of course, is just what a *distribution* does) therefore enables the definition of *fields* of density, temperature, and pressure throughout the region of space occupied by the gas. In intuitive terms, then, a Boltzmann coarse-graining implements a many-to-one map from the exact micro-configuration of the gas to a field description in terms of macroscopic variables like density, pressure, and temperature.

Before moving on, I'd like to underscore an important point about the definition of macrostates. As I have defined the term here, a macrostate does *not* correspond to a specification of density and temperature and pressure and so on for the system *as a whole*—which wouldn't be well-defined outside of equilibrium—but rather to the specification of continuous *fields* of these thermodynamic quantities spread throughout space (or the region of space occupied by the system). Note here the importance of the cell-size constraint: the cells must each contain sufficiently many molecules that quantities such as density, temperature, and pressure are well-defined (individual molecules have no temperature!), yet also be small enough that these quantities do not vary substantially between neighbouring cells, so that it makes sense to speak of quasi-continuous fields of these quantities. Under this constraint, macrostates can be defined even for systems out of equilibrium.

2.1.2 Statistical Mechanics of Equilibrium

Macrostates, as defined above, correspond to elements of a suitable Boltzmann coarse-graining, and thus to regions of phase space. We define the *Boltzmann entropy* of a macrostate M as the logarithm of its phase-space volume,

$$S_B(M) = k_B \log \mu(M). \tag{2.1}$$

Here μ represents some natural phase-space measure, which is chosen so as to be preserved under the (distributional) dynamical evolution. In the case of a classical gas evolving under the Hamiltonian equations of motion, we may take $\mu = \mu_L$, the Liouville-Lebesgue measure.

The equilibrium macrostate M_{eq} is defined as the maximum-entropy macrostate, or equivalently, the macrostate with the largest phase-space volume. Equilibrium statistical mechanics is chiefly concerned with maximising S_B under various different constraints, in order to derive the equilibrium distribution for the system (i.e., the set of occupation numbers that the system instantiates in equilibrium).[2] For instance, under the constraint of fixed total particle number ($\sum_i n_i = N$) and fixed total energy ($\sum_i n_i \epsilon_i = U$), we can derive the Maxwell-Boltzmann distribution, $n(\epsilon) = e^{-\epsilon/k_B T}$.

Evidently, there is nothing within equilibrium statistical mechanics that supplies a time direction—indeed, equilibrium statistical mechanics can be carried through without any reference to time whatsoever. Of greater interest to us, therefore, is not the equilibrium distribution per se, but the *process of equilibration*. It is the relentless progress of systems *towards* equilibrium in one time direction (from earlier times to later), and *away* from equilibrium in the other time direction (from later times to earlier) that constitutes the world's time asymmetry, and to which we now turn.

2.1.3 Statistical Mechanics of Non-Equilibrium

What explains why a system, out of equilibrium at one time, will tend towards equilibrium at later times? A common sort of argument goes as follows. The equilibrium macrostate M_{eq} is not only the largest macrostate, but it is in fact much, much larger than the others. Thus, if a phase point moves around the phase space "at random", it must sooner or later find itself wandering into M_{eq}, where it will then stay for a very long time.

While this argument has a true premise (for large N, the equilibrium macrostate is indeed very much larger than the others) and may seem intuitively compelling, there are reasons for dissatisfaction. Most obviously, it is unclear how to understand the suggestion that the phase point moves about "at random". As we know, the phase point in fact moves along a deterministic trajectory given by the Hamiltonian dynamics. It is perfectly conceivable that those dynamics would take the phase point on small loops within the initial non-equilibrium macrostate, or indeed simply take it even further away from equilibrium (indeed, as Frigg (2009) points out, there are perfectly well-defined Hamiltonians that do this). Perhaps we are to

[2] In practice, it is more common to maximize $\log W$, where $W = \frac{N!}{n_1! n_2! \ldots n_m!}$ is called the *multiplicity* of the distribution. If all the cells i have the same size, then W scales linearly with the phase-space volume associated to the distribution, and so there is no discrepancy in the mathematics. Arguably, however, the definition of Boltzmann entropy using phase-space volume is conceptually preferable to the definition involving the multiplicity as it circumvents awkward metaphysical questions about haecceitism; see Albert (2000, pp. 45–47).

suppose that such phase-space trajectories do not occur, or at any rate are very atypical—but if that's the idea, then this premise deserves to be made fully explicit.

Suppose, then, that we grant the following premise:

Equilibration 1 (Forwards)**.** *The "vast majority", in the sense of conditional μ-measure, of the microstates in any given non-equilibrium macrostate M are taken by the dynamical evolution ϕ_t into higher-entropy regions of the phase space and ultimately into M_{eq} as we evolve the system* forwards *in time.*

This premise forms the core of the "neo-Boltzmannian" explanation for equilibration. I assume that it is true. Nevertheless, there is a serious obstacle in the way of *using* this premise to account for the time-asymmetry of the process of equilibration. The problem is that, subject to a few plausible additional assumptions (see Earman (2006)), Equilibration 1 entails the troublesome counterpart:

Equilibration 2 (Backwards)**.** *The "vast majority", in the sense of conditional μ-measure, of the microstates in any given non-equilibrium macrostate M are taken by the dynamical evolution ϕ_{-t} into higher-entropy regions of the phase space and ultimately into M_{eq} as we evolve the system* backwards *in time.*

Thus, suppose we were to come across an isolated system out of equilibrium, such as a thermos flask containing a few half-melted ice-cubes swimming in warm water. If Equilibration 1 can serve as the basis for an argument that the ice cubes will be more melted, and the water cooler, in the *future*, then all else equal it would seem that Equilibration 2 must serve as the basis for a parallel argument that the ice cubes *have been* more melted, and the water cooler, in the *past*. And the latter conclusion, as we know, is almost certainly false; invariably, we find that at earlier times the ice cubes were less melted, and the water warmer—in other words, we find that the system was even *further away* from equilibrium. This clash of Equilibration 2 with what we take ourselves to know of the past is known as the *reversibility objection* to the neo-Boltzmannian explanation for equilibration.

It is descriptively helpful, at this stage, to introduce a new element into the analysis: ensemble probabilities. Suppose the system of ice-cubes in water is in a non-equilibrium macrostate M (i.e., corresponding to half-melted ice cubes in lukewarm water) at some time t_0. And let us imagine defining a uniform (with respect to μ) probability distribution over the phase-space region M. We can then use the Liouville dynamics to evolve this distribution forwards and backwards in time, to obtain probabilities for the macrostate at other times. By a natural extension of the above considerations, it turns out that those probabilities will

be *symmetrical* about t_0: the induced probability distribution over macrostates at $t_0 + \Delta t$ will be *the same* as the distribution at $t_0 - \Delta t$. Supposing that these probabilities give the "correct" predictions for later times $t_+ > t_0$, we must conclude that they get things *completely wrong* for earlier times $t_- < t_0$.[3]

If the uniform-over-the-macrostate probability distribution gets things right at later times but not at earlier times, a natural solution is to define this uniform probability distribution over the *initial* macrostate of the system—that is, the macrostate at the first instant in time at which the system exists. For instance, in the case of the ice cubes in water, let the probability distribution be uniform over the macrostate M_0 at the time when the ice cubes have just been dropped into the water and the system then isolated from its surroundings (in Reichenbach's (1999, pp. 117–118) terms, this constitutes the formation of a "branch system"). The Liouville time-evolution of *this* probability distribution will then make the right predictions about the system for the duration of the system's existence, from its initial creation to its final state of equilibrium. The reversibility objection is overcome, and a time-asymmetry is introduced, by means of a special sort of *boundary condition* at one temporal end of the system's history, the end that lies on the side which we call the past.

The next section explores the consequences of this solution to the reversibility objection when the system under consideration is the universe itself.

2.1.4 The Past Hypothesis

The considerations of the previous section apply, quite generally, to any thermodynamically isolated system. Of special interest, therefore, are their applications to the *entire universe*, which is often treated as an isolated system (though see Earman (2006) for some misgivings about this). In this section, we consider the strange conclusions that result from applying Equilibration 1 and 2, or rather their analogues involving ensemble probabilities, to the phase space of the universe as a whole.

Consider the present macrostate of the universe. It contains a diverse variety of out-of-equilibrium systems, from the melting ice cubes in my coffee cup to the shining stars in the heavens. In particular, it contains what we would ordinarily call *records* of the past: photographs of our younger selves, libraries filled with history books, and fossils in the ground telling of times long before we were born. Additionally, a field of low-temperature radiation—the cosmic microwave background radiation

[3] In this context, we can understand "correct" to mean *either* "in accordance with our well-founded expectations" *or* "in accordance with the observed regularities of the world"; I do not yet commit to any particular interpretation of these ensemble probabilities.

(CMBR)—surrounds the Earth. This provides a glimpse into the early universe, allowing us to look back in time billions of years to the epoch of recombination, the time when it first became favourable for electrons to combine with protons to form neutral hydrogen atoms, and the universe became transparent to light.

Consider now what we expect will happen to all these myriad out-of-equilibrium systems, including the so-called "records", in the future. The ice cubes will all melt away; the photographs and the history books will gradually become yellowed and fragile, and ultimately decompose; the fossils will eventually crumble under the slow forces of erosion; the stars will continue fusing lighter elements into heavier elements, on and on until iron, and then burn out. These are our thermodynamically well-founded expectations for the future. They are also, therefore, what the uniform-over-the-present-macrostate probability distribution would lead us to believe. But because the uniform-over-the-present-macrostate probability distribution induces *symmetric* predictions for the future and the past, we must deduce that this distribution gives overwhelming probability to the possibility that, contrary to all our strongest convictions, virtually all of the present "records" of the past arose as a spontaneous fluctuation from a higher-entropy cosmological state. The fossils, for instance, rather than being *less* eroded in the past, were with overwhelming probability even *more* eroded (and further back in the past they were more eroded still...) Thus, rather than being the remnants of fossilized creatures that lived millions of years ago, we would be led to infer that these creatures never existed; that the precisely sculpted states of rock that we call the fossil record in fact formed over millions or maybe billions of years by mere chance processes of rock grinding on rock.

Let us call those localized macroscopic configurations of the world that *seem* to tell of events in the past "quasi-records". And let us reserve the word "record" for those quasi-records that are *veridical*—i.e., those quasi-records that co-exist in the world together with the events that they seem to record. In these terms, what we learn from the above line of reasoning can be put, more generally, as follows: the uniform-over-the-present-macrostate probability distribution entails that, with overwhelming probability, almost *none* of the quasi-records in the present macrostate are true records (this is all eloquently explained by Albert (2000, pp. 91–93)). That is the extraordinarily strange conclusion that results from applying our schematic results in the previous section to the system of the universe as a whole.

The way out, once again, is to eschew the uniform-over-the-*present*-macrostate probability distribution in favour of the uniform-over-the-*initial*-macrostate distribution. Let us conjecture, then, that the early universe indeed was in whatever low-entropy macrostate the cosmological quasi-records seem to indicate that it was

(a careful discussion of the sense in which the exceptionally smooth early universe was nevertheless indeed in a very low-entropy macrostate is given by Wallace (2010)). And let us postulate, in addition, a uniform probability distribution over that initial low-entropy macrostate. Then we will find, with overwhelming probability, that the entire unfolding of the universe occurs on a monotonically increasing entropy gradient, from the Big Bang all the way through to its ultimate heat death. And just as importantly, we will vindicate with overwhelming probability our conviction that the overwhelming vast majority of the quasi-records that we find all around us are, after all, *bona fide* records.

The conjecture of a low-entropy initial macrostate for the entire universe has been christened by David Albert as the **Past Hypothesis** (PH).[4] I will sometimes refer to PH more specifically as the *Low-Entropy* Past Hypothesis, to distinguish it from the postulate of a functionally simple (e.g., part-wise uniform) initial probability distribution over the universe's phase space, which Wallace (2023) calls the "Simple Past Hypothesis". These two ingredients—a low-entropy initial macrostate for the universe and uniform probability distribution over that macrostate—form the axiomatic core of the Mentaculus, as we'll see in Chapter 3.

2.2 Lewis's Theory of Counterfactuals

For reasons I alluded to in Chapter 1, much contemporary work on the time direction of causation has focussed in the first instance on constructing an account of the *time asymmetry of counterfactual dependence*. We tend to think that, if some aspects of the present were to be different, then various aspects of the future would also need to be different, while the past would remain largely the same. What explains this time asymmetry of counterfactual dependence?

In 1979, David Lewis published a stupendously influential paper in answer to this question. This paper was in part a response to Kit Fine's criticisms of his possible-worlds account of counterfactuals, published six years previously. In outline, Lewis had proposed that the counterfactual "if it were to be the case that A, then

[4]Note that, for Albert, PH indeed has the character of a conjecture, or primitive postulate. Albert argues that we couldn't possible have *evidence* for PH, since all such evidence (e.g., the CMBR) would itself only be legitimate under the presupposition of PH. Wallace argues, by contrast, that it suffices to postulate that the initial *probability distribution* over the universe's phase space is "simple" (e.g., uniform), *without* presupposing anything about the initial macrostate. The latter is then a matter of straightforward inference from the cosmological evidence. I'm not convinced that either assumption has the epistemic status of a primitive postulate, but I shall not take up the issue here. What matters for our purposes, and about which Albert and Wallace agree, is that both ingredients (low-entropy initial macrostate and simple probability distribution over the initial macrostate) are needed to recover a universe at all resembling ours.

it would be the case that C" is true just in case all the possible worlds in which A is true that are most similar to the actual world are also possible worlds in which C is true (Lewis 1973).[5] Fine objected that such a semantics would deliver the intuitively wrong verdicts in a wide range of cases. For example, we are inclined to think that the counterfactual "if Nixon had pressed the nuclear launch button, there would have been a nuclear holocaust" is plausibly true. But this doesn't seem to follow from Lewis's possible-worlds semantics, since the occurrence of the imagined nuclear holocaust makes for a *very* different world from ours, and so presumably the possible worlds in which Nixon presses the button that are most similar to the actual world are ones in which, for whatever reason, this doesn't eventuate in a nuclear holocaust (Fine 1975).

Lewis's 1979 reply pointed out that his similarity-based possible-worlds semantics for counterfactuals is only schematic. The truth-value of any *particular* counterfactual can only be determined once we have laid down an appropriate (and potentially context-sensitive) similarity metric on the space of possible worlds. And Lewis argued that we should evaluate candidate similarity metrics with an eye to securing the right truth-values in the more obvious cases—of which Fine's case is an example. More generally, Lewis observed that the future is generically quite sensitive to counterfactual stipulations about the present whereas (as we tend to think) the past is not. The challenge, as he saw it, is to find a similarity metric that can account for this asymmetry of counterfactual dependence.

Lewis proposed that the overall similarity of a world w to the actual world @ be judged by the weighted contributions of four factors. In order of importance: (1) the degree to which w violates the laws of the actual world in gross, diverse, and widespread ways, (2) the spatio-temporal extent of the region of *perfect match* between w and @; (3) the degree to which w contains small, localized, and comparatively innocuous violations of the actual laws; and finally (4) the extent of the region of approximate match between w and @ (though Lewis says this latter factor might be accorded negligible or even zero weight).

Lewis's contention was that in "worlds like ours" these criteria of similarity, taken in conjunction with his possible-worlds semantics for counterfactuals, suffice to capture the intuitive temporal asymmetry of counterfactual dependence. This is because the contingent structure of our world's distribution of properties and

[5]More precisely, since there may not be any strict boundary to the set of A-worlds, and hence no A-worlds that are strictly *closest* to the actual world, we should say that the counterfactual comes out true just in case there exists a sphere around the actual world containing at least some A-worlds, and within which all the A-worlds are also C-worlds; or, more succinctly, if there is some $A\&C$-world that is closer than *any* $A\&\neg C$-world.

events (i.e., its "Humean mosaic") is such that later times contain many localised determinants of earlier events, whereas the reverse is not the case. Consequently, only small, localized violations of the actual laws are required in possible worlds that *diverge* from the actual world at time t, whereas diverse and widespread law-violations are required in worlds that *converge* to the actual world at t.

To illustrate, consider again Nixon sat in front of the nuclear launch button, at a few moments before (t_-) and a few moments after (t_+) the time at which he (counterfactually) presses it (t_0). At t_-, it is plausible that the determinants of his action are confined to some well-localised and rather subtle events occurring somewhere in his brain, constituting his process of deliberation. Only a small violation of the actual laws of nature (in Lewis's terms, a "small miracle") is required to steer those processes of deliberation onto the doomsday path in which he presses the button. By contrast, at t_+ the world contains a variety of traces and records of Nixon's action at t_0: in Nixon's memory; in the presence (or absence) of an electrical signal running through the wire connected to the launch button; in the light rays receding every more distantly into outer space, bearing the image of Nixon's hand hovering over the button, and so on. In order for the world in which Nixon presses the button at t_0 to reconverge onto the same future as our world (in which he doesn't press the button) we require a diverse and widespread set of law-violations (a "big miracle") to occur in order to erase all these manifold traces of the button-pressing. According to Lewis's ranked similarity criteria, avoidance of big miracles gets priority over achieving this reconvergence onto a common future, and so the button-pressing worlds in which the counterfactual future is the same as that of the actual world are, it turns out, *less* similar to the actual world than the button-pressing worlds in which the future diverges drastically from the actual world.

The chain of argument here is somewhat elaborate, so let's briefly take stock before turning to the objections that can be levelled against it. Lewis grounds the causal asymmetry on the asymmetry of counterfactual dependence, which is in turn grounded on the so-called "asymmetry of miracles"—i.e., the asymmetry in the scope of the violations of the laws of nature needed to bring about the *divergence* of a world from @ compared to that needed to bring about *convergence* of a world onto @. The latter asymmetry finds its source in the (contingent) asymmetry of overdetermination that obtains in our world: the fact that any given time-slice of our world contains many localized records of *earlier* events (or at any rate, quasi-records; see section 2.1.4) but few to no records of later events. Lewis's scheme is notable for its philosophical imaginativeness, systematicity, and aspiration to completeness. And, as I will argue, it also anticipates salient aspects of the statistical-mechanical

approach of David Albert and Barry Loewer—despite Lewis's acknowledgement that he lacks a conception of how statistical mechanics fits into the story!

But despite its merits, Lewis's scheme is vulnerable to a number of damning objections. I will limit myself to describing just two of these, which will be especially illuminating for us in the transition to Albert and Loewer's alternative, statistical-mechanical approach, to be considered next.

2.2.1 First Objection: Lewis's account implies that almost all counterfactuals are false

The first problem is that for ordinary propositions A and C concerning macroscopic states of affairs, it is almost never the case that *all* the closest A-worlds are C-worlds. This is true regardless of how well-supported the counterfactual may be by the familiar macroscopic laws and generalisations that underlie many of our everyday counterfactual judgements. Statistical mechanics tells us that corresponding to any given macroscopic constraint (specified in the antecedent of the counterfactual), there will be at least some compatible microstates that exhibit an anomalous time-evolution, defying our thermodynamically well-founded expectations. In fact, the situation is a little worse than this: statistical-mechanical arguments give us reason to think that in any *neighbourhood* of a given "typical" microstate, exhibiting the expected entropy-increasing time-evolution, we must find at least some "atypical" microstates, exhibiting the anomalous, entropy-decreasing time-evolution.

Consider: "If I were to put my cup of hot tea into the freezer, it would be colder one minute later." Is it true that in all the closest antecedent-worlds, the consequent obtains? Not if by Lewis's "possible worlds" we understand fully-fledged, microscopically detailed possible histories of the world, and if by Lewis's "law-violations" we understand violations of the fundamental dynamical laws governing the evolution of the microstate. For in this case, in any neighbourhood of an $A\&C$ world we will find at least one $A\&\neg C$ world, and so among the closest A-worlds there will inevitably be a few, rare, anti-thermodynamic $\neg C$-worlds in which, in the minute subsequent to placing the tea in the freezer, the net flow of heat is from the freezer into the cup. It should be emphasized that this is a point about the *topological* structure of phase space, and so this objection holds up quite independently of the choice of "closeness" metric.[6] Thus, unlike for Fine's objection, this one cannot be dispensed with by moving over to a different weighting of similarity criteria.

[6]It is analogous to how, in the ϵ-neighbourhood of any given real number, no matter how small we make ϵ, we find at least one (in fact: a countable infinity) of rational numbers.

It might be argued that Lewis can be sanguine about this implication of his analysis. Alan Hájek has put forward general arguments for the conclusion that most counterfactuals are false (Hájek 2020). If Hájek's arguments are sound, then it would be no impugnment of Lewis's view that it renders false such obvious counterfactuals as the one above about my cup of tea. However, I think the consequences of biting this particular bullet are exceedingly dire, especially if we are interested in the prospect of a counterfactual theory of causation—for we should surely hope to avoid the slide into unrestrained scepticism about causal relations! It would be better to find a semantics of counterfactuals that avoids this conclusion.

2.2.2 Second Objection: There is no asymmetry of miracles

As we have seen, Lewis appeals to a contingent "asymmetry of miracles" that obtains in our world in order to explain the asymmetry of counterfactual dependence with his similarity-semantics for possible worlds. Elga (2001) argues that Lewis's account fails because our world doesn't exhibit an asymmetry of miracles. To understand Elga's argument, we must recall the following striking fact (see section 2.1.3): *for any given non-equilibrium macrostate, the "vast majority" of the microstates compatible with that macrostate determine entropy-increasing histories in both temporal directions.* Thus, in particular, for any given non-equilibrium macrostate of the world containing quasi-records of past events, the "vast majority" of the microstates realising those macrostates do not determine histories in which the quasi-recorded events actually occurred. In other words, for the "vast majority" of the compatible microstates, the local macroscopic "determinants" of the past are *spurious*—not true determinants at all.

Now, the actual microstate is, we take it, among one of the very special microstates that do not have this property, for which the entropy decreases into the past and any macroscopic traces of past events can be trusted. But, as Elga tries to make plausible, a modest perturbation of the microstate at t can easily suffice to move the world over onto a more typical microstate, for which the fundamental dynamical laws determine an entropy-increasing trajectory in both temporal directions. This would not, indeed, involve an *erasure* of the local macroscopic features of the world that Lewis points to as determinants of the past, but rather a removal of their status as bona fide determinants. These quasi-records would not be true records.

Consider again Nixon, sitting in front of the nuclear launch button. Lewis tried to argue that while only a small miracle is required at t_- (presumably, somewhere in Nixon's brain) to get Nixon to press the button at t_0, a big miracle would be required at t_+, involving at least the erasure of all the disparate macroscopic

records in the t_+-macrostate of the events at t_0. Elga demurs. While a big miracle would indeed be required to erase all those localized macro-configurations that apparently provide traces of the events at t_0, *only a small miracle is required to remove their status as bona fide traces*. Thus, it is just as easy to shuffle around a few molecules in Nixon's brain *after the fact* to get him (in accordance with the laws) to press the button at t_0 as it is to do so before the fact; the latter merely involves concomitantly falsifying a number of spurious quasi-records. So the *convergence* to the actual world at t_+ after all requires miracles of the same size as *divergence* from the actual world at t_-, and Lewis's account of the basis for the asymmetry of counterfactual dependence fails on its own terms.

One final comment. It should be noted that these two objections are particularly threatening for Lewis when taken together. For it might be thought that the first objection makes a tedious appeal to a mere technicality. Can't Lewis retain the spirit of his account, and just add the caveat that we ignore $A\&\neg C$ worlds when these are exceedingly rare amongst the $A\&C$ worlds? In light of the second objection, the answer is an emphatic No. For it is precisely those exceedingly rare worlds, scattered in and amongst the typical antecedent-permitting worlds that lie at an entropy minimum, that we need to somehow extract for special consideration when evaluating past-directed counterfactual conditionals.

2.2.3 Transition to the Mentaculus

Despite these objections, Lewis's account retains an intuitive appeal. Even if the details do not quite work out as they need to, it might be held that there is something right about the gist of Lewis's story. It just needs to be modified to overcome the problems adumbrated above. These problems, however, seem to point the way forward to their own solution.

A natural way to deal with the first problem, as suggested above, would be to tweak Lewis's semantics for counterfactuals so that the truth of "if A would C" doesn't require that strictly *all* the closest A-worlds are C-worlds. A mildly revisionary proposal might require that merely *almost all* the closest A-worlds are C-worlds, according to some natural class of measures (i.e., a typicality notion). A more revisionary proposal would introduce probability. Under this proposal, "if it were the case that A, then the probability of B would be x" is true just when the conditional probability over the closest A-worlds that are B-worlds equals x, according to some specified probability measure over the set of possible worlds.

On its own, this doesn't yet solve the second problem—indeed, as I've alluded, it potentially makes it worse! But this second problem too can be solved in a way

that is compatible with the revisionary proposals above. The trick is to restrict our consideration solely to those possible worlds that resemble ours in the crucial respect that *almost all quasi-records are true records*. A natural way of incorporating this restriction into Lewis's scheme would be to liberalize the conception of what counts as a law of our world to include not only the *dynamical laws* but also, crucially, the Past Hypothesis (see section 2.1.4).[7] Doing so, we find that convergence of a world to @ once again requires a "big miracle", though of an unusual sort: on top of whatever minor, localized violations of the *dynamical* laws are required at the time of reconvergence, a major violation of the Past Hypothesis would need to be called upon. This would make the divergent world more similar to @ than the reconvergent world, even though both involve a large region of perfect match with @, and both involve comparably significant violations of the dynamical laws of @.

Invoking both of these modifications to Lewis's account gets us to the following view about the semantics of counterfactuals: "if it were the case that A, it would be the case that C" is true just in case *almost all* the closest $A\&PH$-worlds are C-worlds. And this, to a first approximation, coincides fairly well with the account of counterfactual time asymmetry promulgated by David Albert and Barry Loewer, to which we now turn.

[7]For the avoidance of confusion: Whether a particular quasi-record constitutes a true record is *logically* independent of the Past Hypothesis. However, conditional on the Past Hypothesis it becomes overwhelmingly *probable* that quasi-records are true records, at least during the early cosmological epoch in which we find ourselves. Moreover, in worlds like ours in which most quasi-records seem to tell of a low-entropy past, the failure of the Past Hypothesis entails the failure of the principle that almost all quasi-records are true records.

3
The Mentaculus

3.1 Axioms and Basic Features

Albert & Loewer's account of the counterfactual time asymmetry is derived from their broader framework for naturalized metaphysics, which they call the Mentaculus. The Mentaculus consists of three core ingredients (Albert 2000; Loewer 2020):

1. (L) The underlying dynamical laws of nature (usually assumed deterministic).

2. (PH) A boundary condition at one temporal end of the universe, specifying some particular, special sort of macrostate for the universe at the initial time.

3. (SP) A probability distribution ρ_0 that is uniform, according to some natural state-space measure, over the boundary condition specified in (PH).

I will refer to these three ingredients as the "axioms of the Mentaculus". From these three axioms, most of what we take ourselves to know about the world—including, in particular, the laws of thermodynamics and other "laws" of the special sciences—is supposed to be (at least in principle!) derivable. Loewer therefore argues that they form a "best system" for our world in the Lewisian sense that they strike the optimal balance between simplicity and informativeness.[1] Together, they form

[1]On a Humean view of laws of nature, it is therefore these, rather than the dynamical laws alone, which qualify as the fundamental laws of nature. Of course, L taken in conjunction with the *exact physical microstate* of the whole universe at any one time suffices to derive *all* the particular facts about our universe whatsoever (i.e., the full Humean mosaic); this claim to lawhood therefore depends on it being considerably less simple to specify the exact microstate for the universe than it is to specify PH and ρ_0. But the status of PH and ρ_0 is not our primary concern here; for more discussion, see Loewer's 2023 and Eric Winsberg's response (Winsberg 2023).

the core of the system; everything else I will say about the Mentaculus is either derived from these axioms (as, for example, the entropy gradient and the branching structure), or constitutes additional theorizing within the framework demarcated by these axioms (as, for example, the Mentaculus theory of counterfactuals).

These three axioms furnish a probability distribution over the set of all propositions concerning macroscopic states of affairs ("macro-propositions"). To compute the probability that A obtains at time t, simply start with the distribution ρ_0 over the initial macrostate (specified in PH), evolve this distribution forwards to the time t using the distributional version of the dynamical laws (specified in L), and then restrict the time-evolved distribution to the macrostate-region M_A. The integral of the resulting distribution over all of the phase space is the desired (unconditional) probability of $A(t)$. Mathematically,

$$\Pr_{\mathcal{M}}(A(t)) = \int P(M_A)L(t)\rho_0 \, \mathrm{d}\mu, \tag{3.1}$$

...where $L(t)$ is the distributional time-evolution operator (in the classical case, the Liouville dynamics) from the time of the Past Hypothesis to the time t, $P(M_A)$ is the restriction of the distribution to the macrostate-region M_A, μ is the phase-space measure, and the integral runs over all of the phase space.

In a natural extension of this idea, we can define joint probabilities of macro-propositions obtaining at different times. The probability that A_1 obtains at t_1, A_2 at t_2, and so on is given by,

$$\Pr_{\mathcal{M}}(A_1(t_1), ..., A_n(t_n)) = \int P(M_{A_n})L(t_n - t_{n-1})...P(M_{A_1})L(t_1)\rho_0 \, \mathrm{d}\mu. \tag{3.2}$$

Note that an ambiguity arises as to the order of the evolution-restriction operations in the integral. Fortunately, this ambiguity is benign, as permuting these operations has no effect on the value of the integral.[2] Equation 3.2 is thus well-defined, since the permutation symmetry on the LHS is matched by the corresponding symmetry on the RHS.

With all possible combinations of joint probabilities defined, so too are the conditional probabilities. From equations 3.1 and 3.2, it follows that the conditional probability of $B(t')$ on $A(t)$ is given by,

$$\Pr_{\mathcal{M}}(B(t')|A(t)) \equiv \frac{\int P(B)L(t'-t)P(A)L(t)\rho_0 \, \mathrm{d}\mu}{\int P(A)L(t)\rho_0 \, \mathrm{d}\mu}. \tag{3.3}$$

[2]Roughly, this is because the integrand serves simply to pick out all those phase-space trajectories that are at A_1 at t_1 *and* at A_2 at t_2 *and*... The invariance of the phase-space measure μ under time-evolution ("Liouville's theorem") then assures us that it does not matter at which time we perform the integration over these selected trajectories.

In words: To evaluate $\Pr_{\mathcal{M}}(B(t')|A(t))$, we first take the distribution over the macroscopic boundary condition specified in PH, and evolve it forward, according to the laws L, to the time t. We throw away that part of the distribution that lies outside the macro-condition A, and renormalize. We then evolve this renormalized distribution (backwards or forwards) to the time t'. The conditional probability $\Pr_{\mathcal{M}}(B(t')|A(t))$ is then equal to the value assigned to the macro-condition B by this renormalised and time-evolved distribution.

I will now make a few comments regarding the Mentaculus conditional probability function $\Pr_{\mathcal{M}}(\cdot|\cdot)$, which has a number of interesting properties.

First, the choice of the Mentaculus axioms was originally motivated by the need to account for the thermodynamic asymmetries in time, and in particular the asymmetry of equilibration. Isolated systems (including, as is usually presumed, the whole universe) evolve uniformly towards equilibrium in one temporal direction (the direction we call the future) and away from equilibrium in the other (the direction we call the past). The familiar Boltzmannian explanation for the entropy increase in the futureward direction is that the "vast majority" (according to the usual measure) of the microstates compatible with any given non-equilibrium macrostate lie on state-space trajectories that increase in entropy towards the future. And the well-known problem with this explanation is that its truth, taken in conjunction with the time-symmetric dynamical laws (and a few other plausible assumptions), entails that an equal-sized "vast majority" of compatible microstates lie on state-space trajectories that increase in entropy towards the past as well.

How does the Mentaculus get around this infamous "reversibility objection"? The key point is that the Statistical Postulate (SP) employed in the Mentaculus does not postulate a uniform probability distribution over the *present* macrostate, on the basis of which past- and future-directed inferences are made. Rather, it postulates a uniform probability distribution over the *initial* macrostate. The requirement of a monotonic entropy gradient for the whole universe then places a constraint on the sort of macrostate that can serve as the boundary condition postulated in PH. The adequacy condition is, of course, that for $t' > t$, $S(M_A) > S(M_B) \Rightarrow Pr(B(t')|A(t)) \approx 0$, where $M_{A(B)}$ is the region of the state-space corresponding to the macrostate $A(B)$, and S is the Boltzmann entropy. The condition is met by choosing the initial macrostate to be one of very low entropy.

Second, in general we may safely assume that even for fully specified macro-conditions A and B (i.e., maximally specific macro*states*), $\Pr_{\mathcal{M}}(B(t')|A(t))$ does not take an extremal value for $t' > t$. Microscopic differences among the states compatible with $A(t)$ will generically get amplified, over time, into macroscopic

differences in the future evolutions of those states. (Chaotic systems provide a vivid illustration of this sort of phenomenon, and indeed most complex, real-world systems are chaotic; see Mitchell (2009, pp. 15–40) for a primer.) Thus, while the microscopic laws can be assumed deterministic, the *macroscopic* evolution of the universe is indeterministic: full specification of the macrostate at a time t does not fix a unique macrostate at a later time t'.

By contrast, for $t' < t$, we do usually find that $\Pr_{\mathcal{M}}(B(t')|A(t))$ takes an extremal value if $A(t)$ is a fully specified macrostate. This is because, in such a case, we would expect $A(t)$ to contain quasi-records of past events, including the occurrence or non-occurrence of $B(t')$, and the tacit conditionalization on PH will ensure that those quasi-records are indeed veridical with overwhelmingly high probability. In other words, experience teaches that $A(t)$ will bear apparent records of either $B(t')$ or $\neg B(t')$; it is very unusual for a macroscopic event to leave absolutely no traces at later times. Then, given the conditionalization on PH, we know that only phase space trajectories that really pass through $B(t')$ at t' will end up, at t, in macrostates that bear apparent records of $B(t')$. So $\Pr_{\mathcal{M}}(B(t')|A(t))$ is either 1 if $A(t)$ contains quasi-records of $B(t')$, and 0 if it contains quasi-records of $\neg B(t')$; and only very rarely does it take on some non-extremal value, in cases where all traces of relevant events at t' have somehow been erased.

In sum, then, the Mentaculus conditional probabilities $\Pr_{\mathcal{M}}(B(t')|A(t))$ entail not only that the macro-evolution of the universe is *indeterministic*, but, more strongly, that it is *branching*. The branching structure of the world according to $\Pr_{\mathcal{M}}(\cdot|\cdot)$ is depicted in figure 3.1.

We are now in a position to explain, somewhat informally, how the Mentaculus supports a general asymmetry of counterfactual dependence. Consider the counterfactual $A(t) \boxrightarrow B(t')$. Entertaining the antecedent of the counterfactual involves supposing ourselves to be, at the time t, on a different "branch" of the structure depicted in figure 3.1 from the mainline. Such a supposition will generically engender many possibilities for the future macro-development of this counterfactual possible world, all of which differ from the actual future macrohistory. By contrast, on the counterfactual supposition $A(t)$ there will generally only be one possible macrohistory to the past of t, and this must converge, sooner or later, onto the past macrohistory of the actual world. Thus, if it had been the case that $A(t)$, much of the macroscopic history of the world *prior* to t would have matched that of the actual world—all of it, in fact, that lies prior to the time at which the $A(t)$-world (or worlds) branched from the actual world. By contrast, if it had been the case

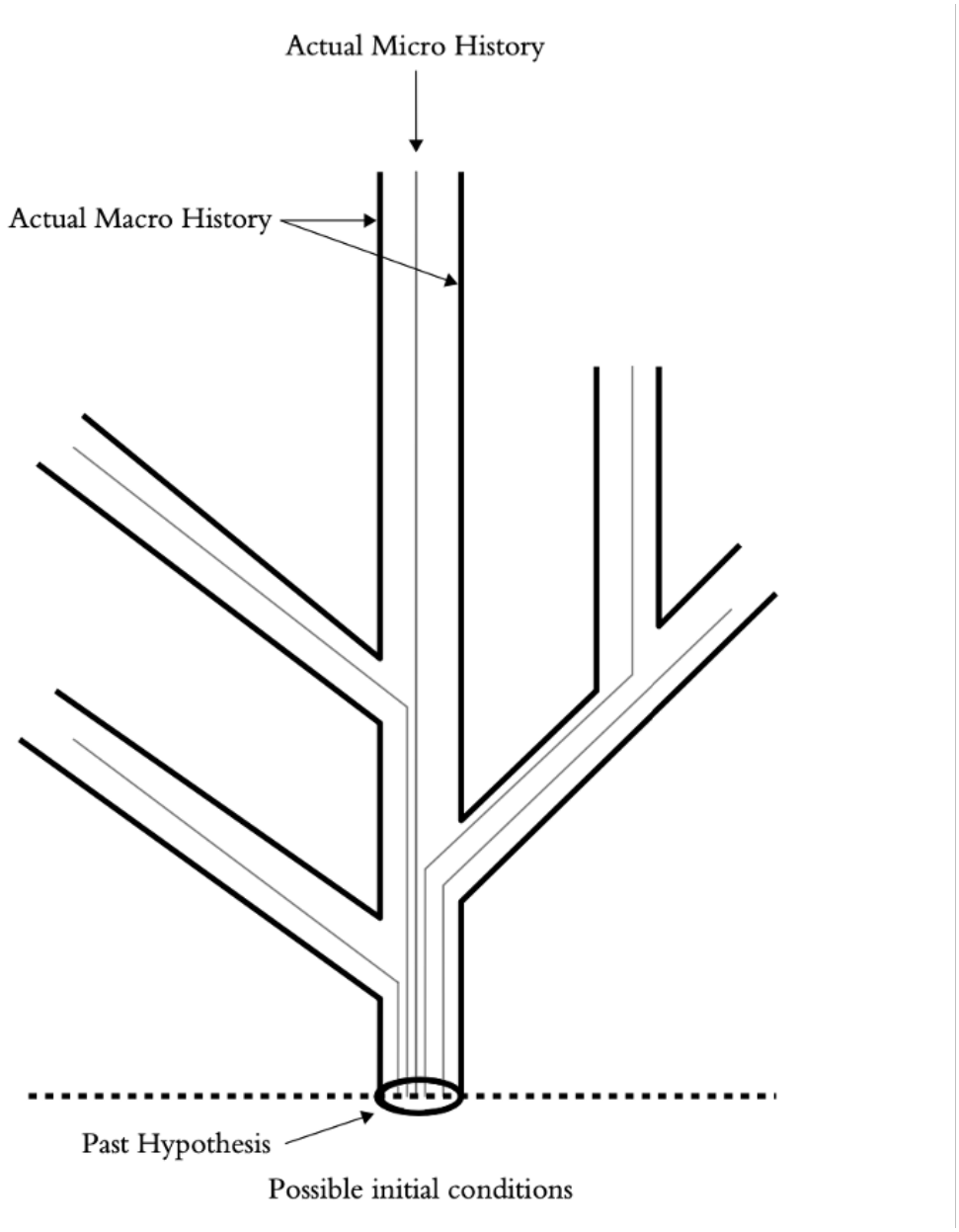

Figure 3.1: The branching structure of the Mentaculus. The macro-evolution of the world according to $\Pr_{\mathcal{M}}(\cdot|\cdot)$ is indeterministic into the future but deterministic into the past. Figure taken from Loewer (2006).

that $A(t)$, then of the numerous (macro-)futures open to the world *after* t, precisely none of them coincide with the numerous futures open to the actual world after t.

In the remainder of this chapter we will shape these informal ideas into something more precise, defining a formal semantics for counterfactuals based on the Mentaculus conditional probability function $\Pr_{\mathcal{M}}(\cdot|\cdot)$, before explaining how this probability function supports forwards-predictive macrodynamical equations. Both of these ideas will be crucial for the discussion of causal models in Chapters 4 and 5.

3.2 Deliverances of the Mentaculus

3.2.1 Theory of Counterfactuals

In section 3.1, I explained how the Mentaculus conditional probabilities $\Pr_{\mathcal{M}}(\cdot|\cdot)$ induce a branching structure for the macroscopic development of the world, and I gave a rough, qualitative argument to the effect that this branching structure can be expected to give rise to a generic time asymmetry of counterfactual dependence. This argument went as follows. Suppose $A(t)$ is a contrary-to-fact supposition about the time t. Then, due to the branching structure, there would be many possible macro-futures for the world after time t, all of which generically differ from the actual macro-future. By contrast, in this idealization of perfect branching structure, there would be only one possible macro-history for the world before time t, and moreover this counterfactual macro-history must ultimately converge with that of the actual world. If this convergence happens fast enough, then much of the past would be unchanged under the supposition of $A(t)$. This argument appeals to an implicit semantics for evaluating counterfactuals using the machinery of the Mentaculus. This section is concerned with making that semantics explicit.

Before doing that, however, a small complication must be introduced. The problem with the above argument is that it really only holds up when $A(t)$ amounts to a complete specification of the entire macro*state* at t. But more generally, $A(t)$ can be any macro-proposition we like about how things could have been at t. If, as in the typical case, $A(t)$ is an incomplete specification of the macrostate at t, then there will be several distinct "branches" of the Mentaculus on which $A(t)$ is satisfied. To evaluate the consequences of the counterfactual supposition, which of these branches do we consider? Naturally, and in the spirit of Lewis, we need to consider the "closest" $A(t)$ branches to the actual world. In defining explicit truth conditions for counterfactuals, then, we need think about how to characterise the notion of "closeness" within the confines of the Mentaculus framework.

Loewer (2023) suggests the following:

Counterfactuals 1 (Loewer)**.** *$A(t) \boxdot\!\!\rightarrow P(B(t')) = x$ is true iff there is a time t^* between the time PH holds and t when the macro state is $M(t^*)$ such that $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = x$ and t^* is the time closest to t between PH and t for which $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ is sufficiently large.*

(See also Loewer (2006) for an earlier, very similar proposal, couched in terms of "decisions".) It bears emphasising, once again, how this can be viewed as very much in the spirit of Lewis's semantics. Heuristically, to evaluate "if $A(t)$ would $B(t')$" we look at the conditional probability, according to the Mentaculus, of $B(t')$ conditional on all the $A(t)$ worlds "closest to actuality," where the closest $A(t)$ worlds are all of those reachable, in accordance with the dynamical laws, from the *actual* macrostate at some other time t^*, which is chosen to be as close to t as is nomologically possible while still giving $A(t)$ a "sufficiently large" probability.

One key *difference* between this account and Lewis's is that we have eliminated the need for "miracles". The branching of worlds in the Mentaculus is not due to occasional lapses in the underlying laws of nature (which, at least in classical mechanics, are fully deterministic), but rather due to the fact that many distinct microstates compose each macrostate, and these may come to diverge macroscopically even under their individually lawful time-evolution. In other words, while the *micro*-dynamics is deterministic, the induced *macro*-dynamics is not. By definition, every possible world within the support of the Mentaculus is governed by the same exceptionless micro-physical laws of nature—none of these worlds contain "miracles" in Lewis's sense. Of course, to achieve this property we have also had to eliminate possible worlds that match perfectly at the microphysical level for some times but not others; "perfect match" means for us only perfect match of all macroscopic facts, not perfect match of the exact physical state.

The precise Mentaculus-based semantics for counterfactuals that I favour, however, departs from Loewer's account in a number of respects.

First, Loewer stipulates that t^* lie in between the time of PH and the time t. But this strikes me as unnecessary. If t^* lies "on the other side" of the PH from t (i.e., to the future of t), then the Mentaculus probability $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ will be negligibly small (that is, assuming $A(t)$ is a *contrary-to-fact* proposition about time t). This is due to the records of the actual, *in-accordance-with-fact* events occurring at time t that are present in $M(t^*)$. Consequently, $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ will only be "sufficiently large" if $t^* < t$; the Mentaculus branching structure takes care of this on its own.

Second, Loewer's truth conditions involve the vague phrase "sufficiently large", which he argues is context-sensitive, analogous to the context-sensitivity of the

similarity relation in Lewis's similarity semantics. This too strikes me as an unnecessary complication. The point about the probability $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ being "sufficiently large" is that we would like the counterfactual state to evolve from $M(t^*)$ to $A(t)$ in a way that is in accordance with our well-founded expectations for lawful *macroscopic* time-evolution (i.e., our expectation that the laws of thermodynamics are obeyed). Since the Mentaculus by definition only has support on microscopic trajectories that obey the constraint imposed by the Past Hypothesis, the probability given to anti-thermodynamic time-evolution from $M(t^*)$ to $A(t)$, at least in our early-universe epoch, is utterly negligible—and in the thermodynamic limit of large system size, it is zero. Thus, we can take "sufficiently large" to mean simply non-negligible (or non-zero in the thermodynamic limit), invariantly to context.

Third, the proposal that we should find the time t^* closest to t for which $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ is sufficiently large (or non-negligible) is conceptually appealing as a precisification of "closeness", but also has a number of awkward properties. One problem concerns the evaluation of counterfactual conditionals with *true* antecedents. Presumably, the time t^* closest to t for which $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ is non-negligible, in the case where $A(t)$ is true, is just t itself. But this choice feels somewhat arbitrary in cases where A obtains over a *period* of time. Why not choose t^* to be any of the *other* times at which A is true in the uninterrupted period containing t? Perhaps worse than this, however, is that this definition introduces an inappropriate asymmetry into the evaluation of "$A(t) \boxdot\!\!\rightarrow \Pr(B(t')) = x$" as compared with "$\neg A(t) \boxdot\!\!\rightarrow \Pr(B(t')) = x'$". In general, if $A(t)$ is true, then we will have $t^* = t$ for the former but $t^* < t$ for the latter, and a possible consequence of this will be to introduce a difference between x and x' that is merely a function of the temporal proximity of $M(t^*)$ to the time t' of the consequent, over and above any salient differences between $A(t)$ and $\neg A(t)$ that bear on the probability of $B(t')$.

To resolve this awkwardness, I suggest defining t^* to be the closest time to t for which *both* $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ *and* $\Pr_{\mathcal{M}}(\neg A(t)|M(t^*))$ are non-negligible. Call this time the *branch-point time* for $A(t)$. This definition solves the first problem because, even if A is true over an extended *period* of time containing t, there is presumably a latest *moment* of time (or something approaching a moment) before that period at which the world could instead have evolved to $\neg A(t)$. It solves the second problem because this definition is now symmetric in $A(t)$ and $\neg A(t)$.

Fourth and finally, in keeping with Lewis's account, we would like the counterfactual $A(t) \boxdot\!\!\rightarrow \Pr(B) = x$ to come out *vacuously true* if there is *no* closest $A(t)$ world—that is, if there is no time t^* for which $\Pr_{\mathcal{M}}(A(t)|M(t^*))$ is non-negligible.

Assimilating the lessons from these observations, I propose the following modified version of Loewer's Mentaculus-based semantics for counterfactuals:

Counterfactuals 2. *$A(t) \boxdot\!\!\rightarrow P(B(t')) = x$ is true iff $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = x$, where t^* is the branch-point time for $A(t)$ (or there is no such time t^*).*

3.2.2 Macrodynamics Predictive Forwards in Time

If the Mentaculus is the "best system of the world", then there is a sense in which the Mentaculus conditional probabilities $\Pr_{\mathcal{M}}(\cdot|\cdot)$ are indeed the "correct" ones. But it is patently a hopeless enterprise to try to explicitly *use* those probabilities to predict the macroscopic development of the world (or rather: whatever small parts of the world are of interest). Even if the underlying dynamical laws were perfectly well understood, and even if the initial macrostate of the universe was perfectly known, it would not be even remotely computationally feasible to actually use those dynamical laws (in their distributional form) to evolve the uniform-over-the-PH distribution on the phase space of the universe forwards to the present time. Clearly, some sort of approximation scheme must be used instead.

In *Time and Chance*, Albert discusses one such approximation scheme. Assume, once again, that we are interested in the time evolution of some set of macro-properties of a system (closed for all practical purposes), which as usual correspond to regions of the system's phase space. Then, Albert suggests, when considering the macroscopic time-evolution of the system from an earlier time t_1 to a later time t_2, start with the probability distribution that is *uniform* (with respect to the natural measure on the system's phase space) over the macrostate of the system at the earlier time t_1, and use the dynamical laws to evolve this distribution forwards to the time t_2 (Albert 2000, p. 66). (N.B., This uniform-over-the-present-macrostate distribution, of course, generically looks quite different from the distribution over the system's phase space that is induced by the cosmological Mentaculus probabilities!) So, for instance, to calculate the probability of $B(t')$ conditional on $A(t)$, we would start with the uniform probability density over A (which we'll denote ρ_A), and evolve *that* forwards by a time $t' - t$, i.e.,

$$\Pr(B(t')|A(t)) = \int P(B)L(t' - t)\rho_A \, \mathrm{d}\mu. \tag{3.4}$$

Now, equation (3.4) can be related to the Mentaculus probabilities (3.3) by the interposition of a *Gibbs coarse-graining* operation C, which takes as input an arbitrary probability density distribution with support in macrostates $M_1, M_2, ..., M_k$, and outputs the unique distribution that is part-wise uniform over $M_1, M_2, ..., M_k$ while

leaving the probability assigned to each macrostate invariant.[3] In other words, within each macrostate-region separately, C redistributes the associated probability mass of the density distribution uniformly over the region. In this case, since $P(A)L(t)\rho_0$ only has support in the phase-space region A, and assuming that A corresponds to a fully specified macro*state* (rather than a union of many macrostates), the action of C is simply to replace this distribution with the one that is uniform over A and zero elsewhere. Thus, we can write ρ_A as $\frac{CP(A)L(t)\rho_0}{\int CP(A)L(t)\rho_0 \, \mathrm{d}\mu}$,[4] and (3.4) becomes:

$$\Pr(B(t')|A(t)) = \frac{\int P(B)L(t'-t)CP(A)L(t)\rho_0 \, \mathrm{d}\mu}{\int CP(A)L(t)\rho_0 \, \mathrm{d}\mu}. \tag{3.5}$$

Consequently, if the Mentaculus supplies the "true" underlying probabilities, which manifest in the world's empirical regularities, then from the structure of 3.5 we observe that Albert's approximation scheme (equation 3.4) is empirically adequate only provided that the interposition of the coarse-graining operator C leaves the value of the Mentaculus probability (approximately) invariant. Before elaborating on this comment, I must first make two refinements to this general scheme.

First, as Wallace (2023) points out, the coarse-graining operation $C : \mathcal{D} \to \mathcal{D}$ need not be—and in practice rarely is—the exemplar rule described above, in which each equivalence class of phase-space probability density distributions (defined with reference to the associated probability distribution over *macrostates*) get mapped onto a chosen "exemplar"—in this case, the distribution within each equivalence class that is uniform across each macrostate region. Instead, Wallace argues, a variety of coarse-graining operations are available, of which the particular exemplar rule described above is but one. All coarse-graining operations are however subject to the following conditions:

1. For all distributions ρ, $C^2(\rho) = C(\rho)$
2. For all distributions ρ, $CR(\rho) = RC(\rho)$
3. $\int_M C(\rho) = \int_M \rho$, for any macroproperty M

[3]See Wallace (2018). Note that this notion of a *Gibbs* coarse-graining is quite different from the *Boltzmann* coarse-graining defined in section 2.1.1. While the latter constitutes a map from a point in phase space (i.e., a microstate) to the region of phase space to which it belongs (i.e., its associated macrostate), the former constitutes a map from a rather complicated *probability distribution* over the whole phase space to another, simpler distribution over the phase space.

[4]By a slight abuse of notation, expressions involving ρ_0 (such as "$P(A)L(t)\rho_0$") are here to be interpreted as the Mentaculus-*induced* probability distribution over the *system's* phase space specifically, rather than the phase-space of the whole universe. This move might require an assumption to the effect that the system's degrees of freedom are (approximately) uncorrelated with the remaining degrees of freedom of the universe at large, or at any rate that they can be treated as such for all practical purposes.

4. $S_G(C\rho) \geq S_G(\rho)$, where $S_G(\rho) = -\int \rho \ln \rho$

The first condition says that C is a projective map—the coarse-graining of a coarse-graining is just itself. The second condition says that C commutes with the time-reversal operation R, defined by $\phi_t(x) = \phi_{-t}(Rx)$ for all microstates x and times t (i.e., intuitively, the operation that maps each phase point to a corresponding "time-reverse" phase point undergoing the reversed time-evolution). Of particular interest to us, however, are conditions (3) and (4). (3) says that the coarse-graining map leaves the probabilities assigned to all the macroproperties invariant. A necessary and sufficient condition is that it leave the probabilities assigned to all the different macro*states* invariant, as above. (4) says that the coarse-graining map increases the functional S_G, known as the *Gibbs-entropy*. The Gibbs-entropy is a natural measure of the *amount of information* about the microstate contained in the distribution ρ: highly uniform distributions, representing complete uncertainty, have a high Gibbs entropy, whereas sharply peaked distributions, representing extreme confidence, have a low Gibbs entropy.

Note that the S_G-increasing property (4) together with condition (3) provide an illuminating qualitative description of what the coarse-graining is doing: it is discarding information about the microstate *without* changing the probabilities of any macroproperties. If a coarse-graining operation of this kind could be interposed (in the manner of equation (3.5)) without loss in predictive power, this would tell us (very roughly) that only the *macroscopic facts*, rather than their microscopic realisation, are relevant to the determination of system's macro-evolution.

Second, a moment's thought makes it clear that even equation (3.4) is intractable to evaluate as written, at least in the cases of interest to us. This is because it still requires us to work with the *microdynamical* equations of motion (represented by $L(t' - t)$), which is (again) patently infeasible when considering thermodynamically large systems. In practice, a common ansatz is to assume that some convenient coarse-graining operation can be applied very frequently throughout the system's evolution, and thus to derive *macrodynamical* equations of motion.

Wallace glosses this method, with a fair deal of generality, in terms of the notion of a coarse-grained *forwards dynamics* induced by C. This is schematized as the dynamics generated by interspersing the (distributional) microphysical dynamical evolution with the coarse-graining operation at very frequent intervals Δt, i.e., $L^{C+}(t)\rho = CL(\Delta t)C...CL(\Delta t)\rho$. Note that while the Gibbs entropy $S_G(\rho)$ is preserved under the microphysical time-evolution L, it is merely non-decreasing

(and generically, increasing) under the forwards-dynamics L^{C+} because each coarse-graining operation is itself Gibbs-entropy increasing unless the prior distribution is already at a Gibbs-entropy maximum (see Wallace (2023); see also his (2015)).

A concrete example may help to bring this schema to life. Consider Boltzmann's derivation of his infamous "H-theorem". Boltzmann was concerned with the statistical model of a gas of N of molecules inside a fixed container with perfectly elastic walls. He modelled this gas by means of a distribution function $f(\vec{r}, \vec{v}, t)$ for the molecules in the gas, specifying the probability density for finding an arbitrary molecule at position $\vec{r}$ and velocity $\vec{v}$ at time t. Boltzmann derived his *transport equation* to describe the evolution of this distribution function over time. His derivation involves an assumption known as the *Stosszahlansatz* (SZA), to the effect that molecules entering into collision are statistically uncorrelated in their velocity (i.e., their joint distribution function factorizes). Boltzmann then defined the H-functional, $H[f_t] = \int f(\vec{r}, \vec{v}, t) \ln f(\vec{r}, \vec{v}, t) d^3r d^3v$. From the transport equation (and so, crucially, from the assumption that the SZA holds at all times) it follows that $dH/dt \leq 0$, where equality obtains when the gas reaches equilibrium at the Maxwell-Boltzmann distribution, $f(\vec{r}, \vec{v}, t) \propto \exp\left(-\frac{m|\vec{v}|^2}{2k_B T}\right)$, which is independent of position and time. This, of course, is Boltzmann's famous H-theorem (Brown et al. 2009).

Now, the SZA can be regarded as performing a coarse-graining operation of precisely the kind characterised by Wallace. It is widely recognised that the SZA cannot be literally *true* at every time. Rather, the SZA is a serviceable device for predicting the future time-evolution of the gas, which is predictively successful at the macroscopic level of description for (first qualification) typical initial microstates, at least on (second qualification) ordinary timescales long before the recurrence time of the gas. One way to think about it is that the SZA instructs us to *discard*, for predictive purposes, whatever correlations may have developed amongst the molecules in the gas over the course of their interactions. Discarding these correlations is equivalent to coarse-graining the "true" joint probability density $f_N(\vec{r}_1, \vec{v}_1; \vec{r}_2, \vec{v}_2; ...; \vec{r}_N, \vec{v}_N; t)$ for the N gas molecules to the probability density given by the product of the marginals, $Cf_N(\vec{r}_1, \vec{v}_1; \vec{r}_2, \vec{v}_2; ..., \vec{r}_N, \vec{v}_N; t) = f_1(\vec{r}_1, \vec{v}_1, t) f_1(\vec{r}_2, \vec{v}_2, t) ... f_1(\vec{r}_N, \vec{v}_N, t)$ (for more discussion of this correlation-discard method, see Sklar (1993, pp. 207–210)).

f_N is of course equivalent to the true ensemble distribution ρ on the phase space. Since the coarse-grained distribution, $f_1 \times f_1 ... \times f_1$, is obtained from f_N by discarding information about inter-particle correlations, the former consequently has lower Gibbs entropy than the latter (it carries less information about the microstate). And indeed, the stipulation that the SZA is to hold at all times amounts to implementing

this correlation-discard coarse-graining repeatedly, as described above. So we see the key elements of Wallace's schema illustrated in this example.

Incorporating these refinements to Albert's picture does not, however, vitiate the key conceptual insights we get from Albert. Given a coarse-grained forwards dynamics induced by C, we can define the (forwards-directed) macrodynamical probability of $B(t')$ conditional on $A(t)$ as:

$$\Pr_{C+}(B(t')|A(t)) = \frac{\int P(B)L^{C+}(t'-t)P(A)L^{C+}(t)\rho_0 \, \mathrm{d}\mu}{\int P(A)L^{C+}(t)\rho_0 \, \mathrm{d}\mu}. \tag{3.6}$$

Insofar as the forwards-dynamics is predictively successful, and under the assumption of the underlying Mentaculus probabilities, we must again conclude that the interposition of coarse-graining operations into equation (3.3)—indeed, *any* of a broader class of such operations, at much more *frequent* intervals—leaves the values of those conditional probabilities invariant. That is the explanation, according to the Mentaculus, for the success of macrodynamics in predicting the future.

In considering the macrodynamical approximation to the Mentaculus probabilities, we have been restricting our attention to the case where $t' > t$—i.e., the case of forwards-directed conditional probabilities. Why this restriction? Does not an analogous procedure allow us to construct a macrodynamical approximation to the *backwards*-directed conditional probabilities of the Mentaculus?

Of course, it is perfectly possible to define an analogous *backwards dynamics* L^{C-} by successively coarse-graining and evolving a distribution backwards in time, $L^{C-}(-t)\rho = CL(-\Delta t)C...CL(-\Delta t)\rho$. But observe that since the coarse-graining operator C is S_G-increasing, any trajectory of the backwards dynamics must also be S_G-increasing. The forwards and backwards dynamics therefore make incompatible predictions: evolving a distribution forwards by a time t using the forwards dynamics, and then backwards by the same time t using the backwards dynamics, is Gibbs-entropy increasing all the way, and so cannot return us to our starting point. And so, as has been emphasized many times over by Albert and others, the predictive success of the forwards dynamics entails a total predictive failure of the backwards dynamics.

To put the point most conspicuously, let us write

$$L^C(t'-t) = \begin{cases} L^{C+}(t'-t), & \text{if } t' > t \\ L^{C-}(t'-t), & \text{if } t' < t. \end{cases} \tag{3.7}$$

We can then define direction-neutral macrodynamical probabilities (with respect to a coarse-graining C) as follows:

$$\Pr_{C}(B(t')|A(t)) = \frac{\int P(B)L^{C}(t'-t)P(A)L^{C}(t)\rho_0 \, \mathrm{d}\mu}{\int P(A)L^{C}(t)\rho_0 \, \mathrm{d}\mu}. \tag{3.8}$$

As we have already said, under the assumption of the Mentaculus conditional probabilities, the predictive success of forwards-directed macrodynamics requires that $\Pr_C(B(t')|A(t))$ approximates $\Pr_{\mathcal{M}}(B(t')|A(t))$ very closely when $t' > t$. Now we can add that the predictive *failure* of the *backwards*-dynamics requires that $\Pr_C(B(t')|A(t))$ *diverge* from $\Pr_{\mathcal{M}}(B(t')|A(t))$ when $t' < t$.

Whence the asymmetry between the success of prediction and the failure of retrodiction? According to the *Simple Dynamical Conjecture* (SDC), the time-evolution of sufficiently *simple* distributions are "forwards predictable" by a coarse-grained macrodynamics (Wallace 2023). By this is meant that one gets the same results by evolving a simple distribution through time *exactly* using $L(t' - t)$ and then coarse-graining the result as one gets by evolving that distribution through time using the coarse-grained dynamics $L^C(t' - t)$. If this conjecture is right, then we can understand the predictive success of macrodynamics *forwards* in time as resulting from the fact that the Mentaculus probabilities are all generated from the exact time evolution of a simple distribution ρ_0 obtaining at the initial time.[5] Backwards macrodynamics, by contrast, involves applying the coarse-grained dynamics to whatever underlying Mentaculus distribution happens to obtain at the time of the attempted retrodiction, which need not be at all "simple". In this case, SDC does not give us grounds to expect the results of the coarse-grained dynamics to coincide with the exact dynamics. Thus, the Mentaculus (in conjunction with SDC) offers an explanation for the impressive success of macrodynamics at *prediction* and its total failure at *retrodiction*, namely the fact that the "true" (i.e., Mentaculus) probabilities are derived from a simple distribution obtaining at the very earliest time (i.e., the time of the Past Hypothesis).

The purpose of this lengthy detour into the predictive success of macrodynamics will become apparent in Chapter 5, where we will use the results of this section to shed light on the statistical-mechanical underpinnings of the so-called "Causal Markov Condition", which establishes a connection between causal relationships and patterns of statistical independence. Before that, however, we must properly introduce the formal setting in which such connections can be precisely stated and carefully explored—the framework of causal models.

[5]Note that the entire integrand of equation 3.6) is one long application of the coarse-grained forwards dynamics to the simple distribution ρ_0.

4
Causal Models

In contemporary statistics, it is common to formally articulate causal assumptions using the machinery of *structural causal models* (SCMs) and their associated *causal graphs*. These causal graphs are *directed*, in the sense that all the edges between nodes in the graph are directed. Taken in conjunction with data, an SCM and/or its causal graph support various powerful methods of causal inference, such as the computation of the causal effect of one variable on another, and the evaluation of particular counterfactual claims. As noted in Chapter 1, these methods of causal inference are highly sensitive to the *directions* of edges in the graph; differently directed edges license different, often contradictory causal inferences. It is therefore natural to characterise the direction of causation, in a time-neutral manner, in terms of the validity or otherwise of the causal inferences that one is permitted to make on the basis of those direction assumptions.

This, then, will be our strategy. In this chapter, I first provide an overview of SCMs, and describe the various causal and statistical inferences that they facilitate (my presentation here largely follows Pearl et al. (2016)). SCMs, it will be noted, support only the evaluation of *definite* counterfactuals and (epistemic) probabilities of such counterfactuals. In order to handle genuinely probabilistic causality, I therefore introduce a generalized formalism in which the deterministic structural equations are replaced by indeterministic transition functions. Using this formalism, I then offer a formal statement of the *intervention criterion*, which relates causation and counterfactuals. Finally, I put these ideas together with the Mentaculus, thereby formulating a naturalistic definition of the conditions under which particular causal relationships obtain.

4.1 Introduction to Causal Inference

4.1.1 Structural Causal Models

A structural causal model (SCM) is a model of the form $(\mathbf{U}, \mathbf{V}, F)$, where $\mathbf{U}$ is a set of so-called *exogenous* variables, $\mathbf{V}$ is a set of *endogenous* variables, and F is a set of functions $\{f_X | X \in \mathbf{V}\}$ that determine how each endogenous variable depends on (some of) the other variables. The exogenous variables do not depend on any other variables for their values. A list of all the endogenous variables written as functions of the other variables on which they depend constitutes a system of *structural equations*. For this investigation, we stipulate that each exogenous variable appears in the structural equation of exactly *one* endogenous variable. Corresponding to each SCM there is a *causal graph*, which represents at a qualitative level the patterns of functional dependency among the variables. Each variable is represented as a node in the graph, and an edge from a node X to a node Y represents the fact that X appears in the structural equation for Y. All edges in the graph are thus *directed*, signifying the difference between the role of a variable as an independent variable within the structural equation for another variable and the role of that variable as a dependent variable within its own equation. It is also generally assumed that causal graphs are *acyclic*, a feature that we will return to in Chapter 5.

As an example, consider the causal graph of figure 4.1. Here $\mathbf{U} = \{U_X, U_Y, U_Z\}$, and $\mathbf{V} = \{X, Y, Z\}$. In the figure I have represented the exogenous variables with square nodes and the endogenous variables with circular nodes for visual clarity, though strictly this is of course superfluous. If we assume that the structural equations are linear, we can write them as

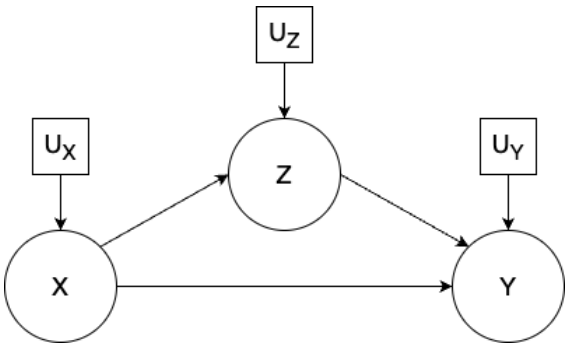

Figure 4.1: A simple causal graph.

$$X = U_X, \quad Y = aX + cZ + U_Y, \quad Z = bX + U_Z, \tag{4.1}$$

for some set of coefficients a, b, c.[1]

[1]These coefficients could be obtained by means of a linear regression on a dataset involving the variables X, Y, and Z. But note that this is only possible under the presupposition of this graphical form for the causal graph, which (as we will soon see) cannot be inferred from the dataset alone. Note also that in general the structural equations need not be linear.

It is important to recognise that the syntactic form of the structural equations is significant in itself, going beyond the significance of the asserted algebraic relationships. Substituting the equation for X into the equation for Z, and the equations for X and Z into the equation for Y, we could write an algebraically equivalent set of equations as

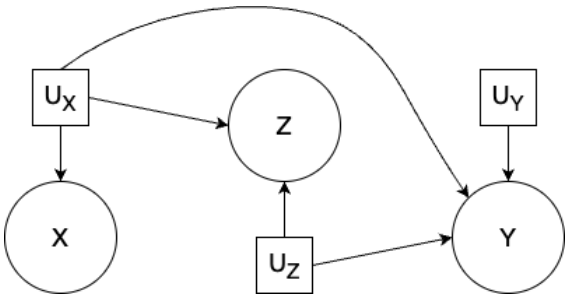

Figure 4.2: A different causal graph.

$$X = U_X, \quad Y = eU_X + cU_Z + U_Y, \quad Z = bU_X + U_Z, \tag{4.2}$$

where $e = a + bc$. Here each endogenous variable is written solely as a function of the exogenous variables U_X, U_Y, U_Z (i.e., in so-called "reduced form"). However, the corresponding causal graph (figure 4.2) looks very different. Since the edges in the graph represent direct causal influence, this means that equations 4.1 receive a different causal interpretation from equations 4.2 (c.f., Woodward (2004, pp. 330–332).

Probabilities are introduced by means of a probability distribution $P(\mathbf{u})$ over the domain of the exogenous variable set $\mathbf{U}$. (Here "$\mathbf{u}$" represents an n-tuple of values for all the exogenous variables.) Since by definition the exogenous variables are all causally independent of each other, then, assuming that causal independence entails probabilistic independence, the distribution $P(\mathbf{u})$ should always factorize, e.g., $P(u_X, u_Y, u_Z) = P(u_X)P(u_Y)P(u_Z)$.

The distribution $P(\mathbf{u})$ induces, via the structural equations, a probability distribution over the endogenous variables. That is, for any subset $\mathbf{\Sigma} \subseteq \mathbf{V}$ of the endogenous variables, we can define (c.f., Pearl (2009, p. 205))

$$P(\mathbf{\Sigma} = \boldsymbol{\sigma}) = \sum_{\{\mathbf{u}|\mathbf{\Sigma}(\mathbf{u})=\boldsymbol{\sigma}\}} P(\mathbf{u}). \tag{4.3}$$

Conditional probabilities can then be defined from the above joint probabilities in the usual way, using the definition $P(A|B) =_{df} P(A\&B)/P(B)$.

4.1.2 The Rule of Product Decomposition

An important consequence of the definition 4.3 of the probabilities in the SCM formalism, and the probabilistic independence of the exogenous variables, is that the joint distribution over the *endogenous* variables factorizes according to the formula,

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i|\mathbf{pa}_i), \tag{4.4}$$

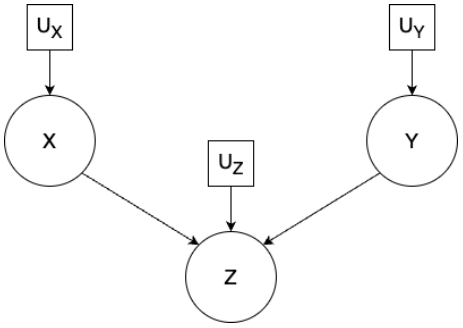

Figure 4.3: Monty Hall problem, with the usual causal structure.

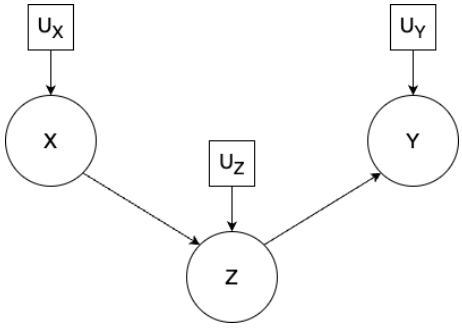

Figure 4.4: Monty Hall problem, with a different causal structure

...where the product runs over all the endogenous variables and $\mathbf{pa}_i$ stands for the values of $\mathbf{PA}_i$, the *parents* of X_i (i.e., the variables taken as arguments in the structural function for X_i). This is known as the *rule of product decomposition*. This formula represents a constraint imposed by the causal graph on the form of the joint probability distribution over the variables (Pearl et al. 2016, p. 29).

To illustrate, consider the two causal graphs in figures 4.3, 4.4.

Figure 4.3 depicts the causal graph for the classical Monty Hall problem, with X representing the location of the car, Y representing the door selected by the player, and Z representing the door opened by the gameshow host (and each variable taking values in $\{1, 2, 3\}$).[2] In the classical Monty Hall problem, the gameshow host observes the value of X and Y, and always chooses the value of Z to be different from both. Let us assume that the variables X and Y take on each of their values with equal probability, and that Z takes on any of its possible values (i.e., values not equal to X or Y) with equal probability. The player wins if they choose the door with the car behind it—that is, if $X = Y$, irrespective of Z. Thus,

$$P(\text{win}) = P(X = Y) = \sum_{x,z} P(X = x, Y = x, Z = z). \tag{4.5}$$

Using the formula for product decomposition, we have that $P(X = x, Y = x, Z = z) = P(X = x)P(Y = x)P(Z = z|X = x, Y = x)$. Thus we compute,

$$P(\text{win}) = \sum_{x} P(X = x)P(Y = x) \sum_{z} P(Z = z|X = x, Y = x). \tag{4.6}$$

The summation over all values of Z evaluates trivially to unity. The remaining factor evaluates to $3 \times 1/3 \times 1/3 = 1/3$. So the probability of winning if the player sticks with their initial choice is $1/3$ (and of course, more famously, the probability of winning if they *switch* is $2/3$).

[2]The Monty Hall problem is suggested as a practice problem (study question 1.5.4) in Pearl et al. (2016). This treatment of the problem is my own.

Note, however, that if we flip one of the arrows in the causal graph to obtain figure 4.4, in which the host's choice of which door to open causally influences the player's choice rather than vice versa (e.g., because the host already opens a door before the player has made their choice), then we reach a different conclusion. With this causal graph, product decomposition gives us $P(X = x, Y = y, Z = z) = P(X = x)P(Z = z|X = x)P(Y = y|Z = z)$. Thus:

$$P(\text{win}) = \sum_{x,z} P(X = x)P(Z = z|X = x)P(Y = x|Z = z), \tag{4.7}$$

...which is easily seen to evaluate to 1/2 rather than 1/3. Over many iterations of the game, this difference in the calculated probability $P(\text{win})$ will show up in the relative frequency with which players win the game. Similar arguments rule out reversing the direction of the other arrow, or both simultaneously, or simply dropping a causal dependency entirely. We therefore find that the causal graph for the Monty Hall problem is entirely fixed by the statistical relationships between the variables.

At this point we can once again note that there is, intuitively, another way to work out whether the arrow should go from Y to Z or from Z to Y. Intuitively, we assume that if the gameshow host opens their door only *after* the player has made their first choice, then the arrow must go from Y to Z; and conversely, if the gameshow host opens their door *before* the player has made their first choice, then the arrow must go from Z to Y. In other words, the direction of the arrow must, we presume, point forwards in time. But the statistical criterion elucidated above makes no reference to the time order of Y and Z. Thus, a substantive question arises as to what justifies this inference from the implied causal direction of statistical relationships to the time order of the causally related events.

4.1.3 Interventions and Counterfactuals

The previous section considered a case (the Monty Hall problem) where a purely statistical criterion (viz., the rule of product decomposition) suffices to nail down a unique causal graph for the situation. We already know from Chapter 1 that such a procedure for causal discovery is not possible in general: the dataset in Simpson's paradox (table 1.1) is compatible with multiple non-equivalent causal interpretations. We can now revisit the discussion of the Simpson's paradox case with the added clarity of the causal-models formalism.

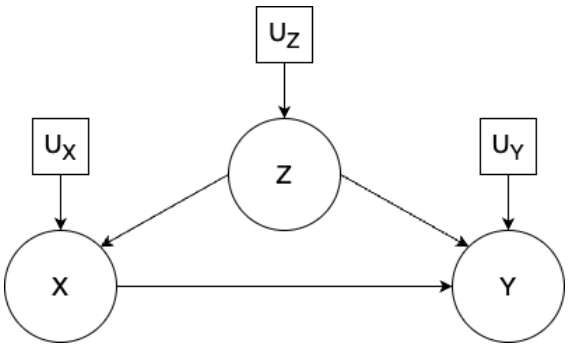

Figure 4.5: An alternative causal model.

Figure 4.1 depicts one possible causal graph for the case, with X representing whether someone takes the drug or not, Z representing whether their blood pressure is high or low, and Y representing whether the person recovers or not. In Chapter 1 we considered the crucial difference between the causal graph of figure 4.1 and the alternative graph that results from reversing the direction of the causal arrow between X and Z, as shown in figure 4.5. Both of these causal graphs, it was said, are compatible with the same dataset (table 1.1). We can now understand this equal compatibility with the data as a simple consequence of the rule of product decomposition, which cannot distinguish the two possibilities. In the first case, product decomposition gives $P(x, y, z) = P(x)P(z|x)P(y|x, z)$, while in the second case, it gives $P(x, y, z) = P(z)P(x|z)P(y|x, z)$. But we have $P(x)P(z|x) \equiv P(z)P(x|z) \equiv P(x, z)$, so these graphs in fact deliver the very same product decomposition formula.

In Chapter 1 we also argued that, while both causal models may give rise to the same *statistical relationships*, they support importantly different *causal inferences*, with the model of figure 4.1 licensing an inference to the causal net benefit of taking the drug on recovery, and the model of figure 4.5 licensing an inference to the drug's causal net harm. We must now ask: in virtue of what formal features of the respective causal models is this crucial difference captured?

The difference between merely statistical relationships and causal relationships is formally captured in the framework of SCMs by the distinction between *conditioning* and *intervening*. That is, we distinguish the *conditional probability* of Y given X, written $P(Y = y|X = x)$, from the *causal effect* of X on Y, which is defined as

$$\Pr(Y = y|\text{do}(X = x)) \equiv \Pr_m(Y = y|X = x), \tag{4.8}$$

...where $\Pr_m(Y = y|X = x)$ is the conditional probability of Y on X that obtains in the *manipulated* causal model that results from severing all the edges into X (or equivalently: from replacing the structural equation for X with the equation $X = x$, leaving everything else untouched). Equation 4.8 defines the "do" operator, and the application of this operator to a variable X (as in $\text{do}(X = x)$) is referred to as a (hard) *intervention* on X (Pearl et al. 2016, pp. 53–55).

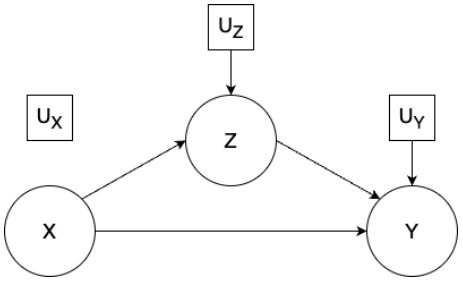

Figure 4.6: Model 4.1 under an intervention on X.

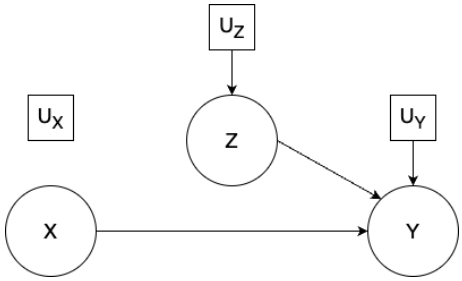

Figure 4.7: Model 4.5 under an intervention on X.

Consider the different effects of an intervention on X in these two causal models. If the causal model of figure 4.1 is the correct one, then applying the intervention on X (by setting $\text{do}(X = x)$) we obtain the manipulated model depicted in figure 4.6. Since all the causal routes by which X may influence Y remain undisturbed through this manipulation, it is easy to convince oneself that $\Pr(Y = y|\text{do}(X = x)) \equiv P_m(Y = y|X = x) = P(Y = y|X = x)$. That is, the causal effect of X on Y is indeed quantified by the conditional probability of Y on X, as manifested in the correlations that obtain between those variables in the population at large.

By contrast, if the causal model of figure 4.5 is the correct one, then an intervention on X yields the manipulated model depicted in figure 4.7. In this manipulated model, the component of the correlation between X and Y that is due to the common cause Z has been broken, leaving only the component of the correlation due to the direct causal influence of X on Y. Thus, we would expect that $P_m(Y = y|X = x) \neq P(Y = y|X = x)$—that is, the intervention disrupts the existing correlation between X and Y. In fact, it can be shown for this case that

$$P_m(Y = y|X = x) = \sum_z P(Y = y|X = x, Z = z)P(Z = z). \tag{4.9}$$

More generally, if we denote by $\mathbf{PA}_X$ ("parents of X") the set of *endogenous direct causes* of X, then the causal effect of an intervention on X can be computed using the formula,

$$P(Y = y|\text{do}(X = x)) = \sum_{\boldsymbol{z}} P(Y = y|X = x, \mathbf{PA}_X = \boldsymbol{z})P(\mathbf{PA}_X = \boldsymbol{z}). \tag{4.10}$$

This is known as the *adjustment formula* (Pearl et al. 2016, p. 57). Note that for the model of figure 4.5, the adjustment formula instructs us to compute the causal effect by looking at the conditional probability of Y on X *and* Z, as would be manifested in the correlations that obtain between X and Y *within each Z-sorted sub-population*. The machinery of SCMs has thus led us to reach the same conclusions by formal means as we reached intuitively in Chapter 1.

The operations of conditioning and intervening can also be applied together, to answer such questions as: "What would be the causal effect of giving someone the drug, given that they have high blood pressure?" These sorts of questions can be answered by means of the following three-step procedure (Pearl et al. (2016, p. 97); see also Pearl (2009, p. 206)):

1. Abduction: Update the probability distribution over the exogenous variables $P(\mathbf{u})$ according to the variables conditioned upon ($\mathbf{\Sigma} = \boldsymbol{\sigma}$), to obtain $P(\mathbf{u}|\boldsymbol{\sigma})$.

2. Intervention: Intervene using $\text{do}(X = x)$ to obtain the resulting manipulated causal model, denoted M_x.

3. Inference: Use M_x, together with the updated distribution over the exogenous variables $P(\mathbf{u}|\boldsymbol{\sigma})$, to compute $P_m(Y = y|X = x, \mathbf{\Sigma} = \boldsymbol{\sigma})$.

Note that for the model of figure 4.5, it does not matter whether we condition on Z and then intervene on X or intervene on X and then condition on Z. That is because, in this case, the structural equation for Z is simply $Z = U_Z$, which does not feature X or any of X's downstream effects. Thus, the conditioning step only updates the probability distribution of U_Z, and updates it in the same way regardless of whether the intervention on X has already been applied or not.

By contrast, for the model of figure 4.1 the order of operations is crucially important. In this model, conditioning on Z *before* intervention updates the probability distribution of U_Z *and* U_X, whereas conditioning on Z *after* intervention updates the probability distribution U_Z only (and moreover updates it in a way that is sensitive to the intervention performed on X). We therefore derive different, incompatible answers to our causal-effect query above depending on the order in which we condition and intervene.[3]

In general, the order of operations is important when either (i) the set of variables conditioned upon intersects the set of variables intervened upon, or (ii) the set of variables conditioned upon intersects the set of effects of any variables intervened upon. But why resolve the discrepancy in favour of the order condition-then-intervene rather than the order intervene-then-condition? There are two compelling reasons for this choice, one philosophical and one practical.

[3]Pearl et al. (2016) calls the former sort of query, where the order of operations is unimportant, "action" queries, and he calls the latter sort, where the order of operations *is* important, "counterfactual queries". But note that this is a narrower usage of the term "counterfactual" than the standard philosophical usage adopted here. As far as we're concerned, *all* computations of causal effects, no matter which variables (if any) are conditioned upon, count as answers to a counterfactual query (i.e., what *would have* happened if we had forced everyone, or everyone in some select sub-population, to take the drug).

The philosophical reason relates to the theory of counterfactuals. Counterfactual queries ask what would have been the case under some hypothetical supposition, holding fixed as many of the actual facts as is possible in consistency with that supposition. By conditioning *first*, we can compute probabilities for the background conditions that *actually* obtained, rather than probabilities for the background conditions that *would have* obtained under the counterfactual supposition. We thereby hold fixed a larger background of actual facts when computing the consequences of the intervention.

The practical reason relates to the scope of the counterfactual queries that may be addressed in this framework. Consider: "What is the probability that someone would recover if they had taken the drug, given that they didn't take the drug?" Intuitively, we would like to calculate something of the form $P(Y = y|\text{do}(X = x'), X = x)$. But if we apply the intervention first, wiping out the structural equation $X = f_X(\mathbf{PA}_X, U_X)$ and replacing it with the equation $X = x'$, then there will be no instances of $X = x$ remaining to conditionalize upon, and the method breaks down. By contrast, we can very well update $P(\mathbf{U} = \mathbf{u}) \rightarrow P(\mathbf{U} = \mathbf{u}|X = x)$, and then intervene on the model $M \rightarrow M_{X=x'}$, without any contradiction arising. The order *condition-then-intervene*, unlike the order *intervene-then-condition*, is therefore more flexible, capable of providing an answer to *any* causal query we like.

Of special philosophical interest is the case of causal effects computed conditional on *all* the endogenous variables. By definition, the structural equations (conceived as a mapping from the exogenous to the endogenous variable set) must have a unique solution $\mathbf{v} = \mathbf{V}(\mathbf{u})$. If the structural equations consist only of invertible functions (as in the case of linear equation systems), then the solution is invertible, $\mathbf{u} = \mathbf{V}^{-1}(\mathbf{v})$. Thus, subject to the assumption of invertible functions, specification of a set of values for all the $\mathbf{V}$ variables gives a unique solution for the values of the $\mathbf{U}$ variables. After intervention, this unique solution $\mathbf{U} = \mathbf{u}$ respectively determines uniquely the values of all endogenous variables in the modified model. In other words, in these circumstances, all causal effects are rendered definite, rather than probabilistic.

When the values of all exogenous variables are given, through the above route or otherwise, the solutions to causal queries take the form of definite counterfactuals, written as $Y_{X=x'}(\mathbf{u}) = y$. This says that in circumstances $\mathbf{u}$, the value of Y under the intervention $X = x'$ is y. It is then possible—and, for us, advisable—to think of the causal effect in its generality as a probability distribution defined on an event space of these definite counterfactuals, arising due to the modeler's *uncertainty* over the details of the actual circumstances (i.e., the value $\mathbf{u}$ of $\mathbf{U}$).

4.1.4 Generalization to Probabilistic Causality

As explained above, the formalism of SCMs supports both the computation of definite counterfactuals, and the computation of probabilities *of* definite counterfactuals arising due to the modeler's uncertainty over the values of the background variables **u**. What it does not support, however, is the computation of counterfactuals with probabilistic consequents. Insofar as we think that some causes may operate not to determine the *value* of a variable, but to determine the *chance* of that variable taking on its different possible values, this is a serious shortcoming of the SCM formalism. Note that here I follow several other authors in taking probabilistic causation to be *deterministic causation of chances* (Hausman 1998, pp. 201–204; Papineau 1989, p. 320; Hausman and Woodward 1999, p. 570).

Indeed, it should have already been clear from the Mentaculus semantics for counterfactuals (Counterfactuals 2) that it will be useful for our purposes to have in hand a generalized formalism in which counterfactuals with probabilistic consequents may be consistently described and computed. However, while several authors have discussed probabilistic causality (Humphreys 1989; Hausman 1998; Woodward 2004), I have been unable to find anyone clearly stating what the models of this formalism would look like, how the methods of causal inference would need to be modified, and how the different sorts of probability in this formalism would interact. I therefore take the opportunity now to fill this gap in the existing literature.

In the formalism of SCMs, the models have the form $\langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{u}) \rangle$, where $F = \{f_i | X_i \in \mathbf{V}\}$ is the set of structural functions, one for each endogenous variable. Each structural function is a map $f_i : U_i \times \mathbf{PA}_i \to X_i$ from the domain of $\mathbf{PA}_i$ (the parents of X_i) and U_i (X_i's exogenous cause), to the domain of X_i itself. As always, $\mathbf{U}$ represents the set of exogenous variables, $\mathbf{V}$ the set of endogenous variables, and I have now also added $P(\mathbf{u})$, the probability distribution over the domain of the exogenous variables.

In the generalized formalism, which I'll call the "indeterministic causal models" (ICM) formalism, we generalize this structure in one place only: instead of specifying deterministic structural functions $f_i : U_i \times \mathbf{PA}_i \to X_i$, ICMs specify a set of *indeterministic* transition functions $T_i : U_i \times \mathbf{PA}_i \to \Delta(X_i)$, where $\Delta(X)$ denotes the set of all probability distributions over the domain of X. Equivalently, and more conveniently, we can write these functions as a map from an enlarged domain into the unit interval, as $T_i : U_i \times \mathbf{PA}_i \times X_i \to [0, 1]$. ICM models then take the form $\langle \mathbf{U}, \mathbf{V}, \mathcal{T}, P(\mathbf{u}) \rangle$, where almost everything is carried over from the corresponding SCM model except that the set F of structural functions has been replaced by a set $\mathcal{T} = \{T_i | X_i \in \mathbf{V}\}$ of indeterministic ("chancy") transition functions.

In the SCM formalism, the structural functions together with $P(\mathbf{u})$ suffice to determine a joint probability distribution over all the endogenous variables (and thus also conditional probabilities, probabilities under interventions, and so forth) via equation 4.3, and this probability distribution automatically satisfies the factorizability condition (equation 5.12). In the ICM formalism, we have no analogue of equation 4.3, but we can instead define a joint probability distribution over the endogenous variables by *insisting* that the factorizability condition be satisfied over all the variables in the model. Thus, we would write,

$$P(x_1, x_2, ..., u_1, u_2, ...) = P(u_1)P(u_2)...P(x_1|u_1, \mathbf{pa}_1)P(x_2|u_2, \mathbf{pa}_2)... \tag{4.11}$$

The $P(u_i)$ are explicitly supplied in the model. If we now identify each of the other factors with the value of the corresponding transition function, $P(x_i|u_i, \mathbf{pa}_i) = T_i(u_i, \mathbf{pa}_i, x_i)$, then all factors are supplied in the model, and so the whole joint probability distribution is defined from the model.

If, as I have been suggesting, we heuristically interpret $P(\mathbf{u})$ as an *uncertainty* over the $\mathbf{U}$ values, and the transition functions T_i as *chances*, then the joint probability distribution over the endogenous variables, given by

$$P(x_1, x_2, ...x_n) = \int P(x_1, x_2, ..., u_1, u_2, ...)\mathrm{d}^n\mathbf{u}, \tag{4.12}$$

generically involves a mixing of chance probability with credence probability. Note, however, that conditional on a given set of values $\mathbf{U} = \mathbf{u}^*$ for the exogenous variables, we obtain from equation 4.11,

$$P(x_1, x_2, ...x_n|\mathbf{u}^*) \equiv P(x_1, x_2, ..., u_1^*, u_2^*, ...)/P(u_1^*, u_2^*...) \tag{4.13}$$

$$= \prod_i P(x_i|u_i^*, \mathbf{pa}_i) \tag{4.14}$$

$$= \prod_i T_i(u_i^*, \mathbf{pa}_i, x_i). \tag{4.15}$$

This leaves us with a joint probability distribution that is wholly chancy, as expected; once the values of the $\mathbf{U}$ variables are given outright, only the chancy transition functions remain to inject probability into the system.

Hard interventions (i.e., of the form $\mathrm{do}(X = x')$) are most naturally implemented by making the replacement,[4]

$$P(x_i|u_i, \mathbf{pa}_i) = T_i(u_i, \mathbf{pa}_i, x_i) \rightarrow P(x_i|u_i, \mathbf{pa}_i) = \delta(x_i - x'). \tag{4.16}$$

[4]Of course, if X is a discrete random variable, then substitute $\delta_{x,x'}$ for $\delta(x - x')$, and replace integrals over the domain of X with summations.

In other words, wipe out the existing transition function for X_i and replace it with a constant (i.e., independent of U_i and $\mathbf{PA}_i$) δ-function peaked on the value to which X_i has been set by the intervention. Causal-effect computations can then be carried out by computing the marginal probability after intervention:

$$P(X_i = x|\text{do}(X_j = x')) = \int \delta(x_j - x') \prod_{k \neq j} P(x_k|\mathbf{pa}_k) \, \text{d}^{n-1}\mathbf{x} \tag{4.17}$$

$$= \int \prod_{k \neq j} P(x_k|\widetilde{\mathbf{pa}}_k) \, \text{d}^{n-2}\mathbf{x}. \tag{4.18}$$

(In the first line, the integral is over all the x_k except for x_i; in the second line, we have integrated over x_j to remove the δ-function, potentially replacing the x_j variable in some of the $\mathbf{pa}_k$ with the fixed value x', yielding $\widetilde{\mathbf{pa}}_k$.)

As a sanity check, note that it should not make a mathematical difference whether we condition or "intervene" on the *exogenous* variables, since in the latter case no graph surgery is required.[5] Mathematically, setting $\text{do}(\mathbf{U} = \mathbf{u}^*)$ involves replacing $P(\mathbf{u}) \rightarrow \delta^{(n)}(\mathbf{u} - \mathbf{u}^*)$. So,

$$P(x_1, x_2, ...|\text{do}(u_1^*, u_2^*, ...)) = \int P(x_1, x_2, ...|u_1, u_2, ...)\delta^{(n)}(\mathbf{u} - \mathbf{u}^*) \, \text{d}^n\mathbf{u} \tag{4.19}$$

$$= P(x_1, x_2, ...|u_1^*, u_2^*, ...). \tag{4.20}$$

This gives us $P(x_1, x_2, ...|\text{do}(u_1^*, u_2^*, ...)) = P(x_1, x_2, ...|u_1^*, u_2^*, ...)$, as expected.

Note from equation 4.15 that even when the exogenous variables are given outright (via conditioning or "intervening"), the joint distribution over the *endogenous* variables remains non-extremal. This, of course, was the whole point of constructing the generalized formalism in the first place. In the ICM formalism, counterfactuals generically take a probabilistic form *even against a fixed background of exogenous variables*. Instead of treating counterfactuals of the form $Y_{X=x'}(\mathbf{u}) = y$, the ICM formalism treats counterfactuals of the form $P(Y(\mathbf{u}) = y)_{X=x'} = p$.

Counterfactuals provide us with the shared reference point needed to locate the formal structure of causal models within the system of the Mentaculus. This we do in section 4.3. But before that, we must first nail down more precisely the connection between causal relationships and counterfactual claims.

[5] Of course, normally we take the exogenous variables to represent unmeasured causes that lie beyond the realm of intervention—uncorrelated "error terms" in our causal models. But since this is merely a check for mathematical consistency, questions of interpretation can be set aside.

4.2 The Intervention Criterion

Before we are in a position to embed the framework of causal models within the more foundational system of the Mentaculus, and thereby explain why causes precede their effects in time, one further piece of set-up is needed. I have several times alluded to the idea that SCMs give us a natural criterion for causal influence in terms of the effects of interventions—what I referred to in Chapter 1 as the *intervention criterion*. It is now time to get precise about just what the intervention criterion says.

In the deterministic setting, the rough idea is that X causes Y if it is possible to manipulate the value of Y by means of interventions on X. We must take care, however, not to rush to an overly simplistic formulation of the relationship between causation and interventions. In particular, although the definition of causal effect is correctly set up insofar as the causal effect of a variable X on a variable Y is non-zero only if X is indeed a cause of Y, the converse is not always true. Consider again the causal model of figure 4.1, in the case where $a = -bc$. In this case, although X acts as a cause of Y both directly and through Z, there is an exact cancellation along the two paths, with the result that the causal effect of X on Y is zero. Interventions on X alone cannot therefore be used to manipulate Y.

We can isolate the direct causal influence of X on Y only if we first intervene to fix Z to one of its possible values. Under an intervention that fixes Z, interventions on X are once again free to produce changes in Y. More generally, in any SCM with variable set $\mathbf{V}$, we can identify a *direct* causal influence of a variable X on another variable Y by first "freezing" with interventions all other variables in $\mathbf{V}$ that lie on a directed causal path from X to Y, and then checking for interventions on X that change the value of Y. With the direct causal relationships identified, it then becomes possible to search for indirect causal relationships by checking for a non-zero causal effect along directed paths through the causal graph; for more discussion, see Woodward (2004).[6]

The above considerations motivate introducing a distinction between types of causal relationships. We will say that X is a *total cause* of Y whenever the causal effect of X on Y is non-zero; we will say that X is a *direct cause* of Y whenever the causal effect of X on Y is non-zero subject to interventions that

[6]The procedure described here for identifying direct causal relationships differs slightly from Woodward's procedure, in that we freeze only those variables in $\mathbf{V}$ that *lie on a causal path* from X to Y, whereas Woodward suggests freezing *all* variables in $\mathbf{V}$ other than X and Y. Remaining strictly within the framework of causal models, both procedures work equally well to identify direct causal influence, and give the same results for the direct causal effect. However, when transplanting these procedures into the Mentaculus, we will see that it is the former and not that latter that gets the right answers.

freeze all endogenous variables lying on a directed path from X to Y; and we will say that X is a *contributing cause* (or simply: a cause) of Y whenever there is a directed path from X to Y, and moreover there is a non-zero causal effect of X on Y subject to interventions that freeze all the endogenous variables not in that path (terminology follows Woodward (2004)).[7]

In all our examples and analysis, the additional subtleties involved in defining contributing causes will not play any important role, so we will focus on total and direct causation. Letting $\mathbf{\Sigma}_{\text{path}}$ denote the set of endogenous variables lying on a directed path from X to Y, and $\boldsymbol{\sigma}_{@}$ denote their *actual* values, we have the following formal intervention criteria.

Intervention Criterion 1 (Total cause)**.** *X is a total cause of Y (in situation $\boldsymbol{u}$) if and only if $Y_{x_1}(\boldsymbol{u}) \neq Y_{x_2}(\boldsymbol{u})$ for some pair of values x_1, x_2 of X.*

Intervention Criterion 2 (Direct cause)**.** *X is a direct cause of Y (in situation $\boldsymbol{u}$) if and only if $Y_{x_1,\boldsymbol{\sigma}_{@}}(\boldsymbol{u}) \neq Y_{x_2,\boldsymbol{\sigma}_{@}}(\boldsymbol{u})$ for some pair of values x_1, x_2 of X.*

A few comments are in order. First, note that $\boldsymbol{\sigma}_{@}$ are the *actual* values of the $\mathbf{\Sigma}_{\text{path}}$ variables, which in the deterministic setting are fully determined by the actual values $\mathbf{u}$ of the exogenous variables.

Second, it might appear that there is a looming circularity in Intervention Criterion 2. A directed path is nothing other than a chain of direct causal relationships; thus, it would seem that we already need to have identified the direct causal relationships in order to apply the criterion for direct causal relationships! In fact, there is no real problem here. We can correctly identify variables lying on a directed path from X to Y by first mapping the full set of *total* causal relationships between the variables in $\mathbf{V}$: certainly, if X is a total cause of Z and Z is a total cause of Y, then Z must lie on a directed path from X to Y. The only risk comes from the failure of the converse condition: X may *fail* to be a total cause of Z (or Z of Y) *even though* Z lies on a directed path from X to Y. But in this case, the directed path running through Z makes no difference to the total causal effect of X on Y, so it can be ignored; if all *other* directed paths are frozen and the total causal effect remains non-zero, then there must be a direct causal connection. In all that follows, we therefore identify the variables in $\mathbf{\Sigma}_{\text{path}}$ by searching for chains of *total* causal influence running from X to Y.

[7]The additional complication in the definition of contributing cause is to rule out possible cases where an intermediary variable Z between X and Y is only sensitive to values of X within a certain range, and Y is only sensitive to values of Z within a certain range, and those ranges are such as to preclude the manipulation of Y along that path by means of interventions on X.

Finally, it will have been noticed that these formal criteria (1, 2) appeal to the idea that if X causes Y, then there should exist two different interventions, $\text{do}(X = x_1)$ and $\text{do}(X = x_2)$, which result in different values of Y. It would seem equally natural to appeal instead to the idea that if X causes Y, then there should be some possible intervention on X, $\text{do}(X = x')$, that results in a value of Y different from its actual value. In other words, why not say instead that X is a total cause of Y if and only if

$$Y_{X=x'}(\mathbf{u}) \neq Y(\mathbf{u}), \tag{4.21}$$

for some possible value x' of X (and analogously for direct causation)? Fortunately, we need not choose: these two precisifications of the notion of total (or direct) cause lead to the same intervention criteria in the deterministic (SCM) setting.[8]

In the indeterministic (ICM) setting, the rough idea is that X causes Y if it is possible to manipulate the *chance* of Y by means of interventions on X. Recall that we are treating probabilistic causation as deterministic causation of chances, so we can try to model the indeterministic intervention criterion on the deterministic one, substituting inequality in the *value* of Y under interventions on X for inequality in the *chance* of Y. Formally, this gives us,

Intervention Criterion 3 (Probabilistic total cause)**.** *X is a probabilistic* total *cause of Y (in situation $\boldsymbol{u}$) if and only if $P(Y(\boldsymbol{u}) = y)_{X=x_1} \neq P(Y(\boldsymbol{u}) = y)_{X=x_2}$, for some pair of values x_1, x_2 of X and some value y of Y.*

Intervention Criterion 4 (Probabilistic direct cause)**.** *X is a probabilistic* direct *cause of Y (in situation $\boldsymbol{u}$) if and only if $P(Y(\boldsymbol{u}) = y)_{X=x_1,\boldsymbol{\Sigma}_{path}=\boldsymbol{\sigma}_@} \neq P(Y(\boldsymbol{u}) = y)_{X=x_2,\boldsymbol{\Sigma}_{path}=\boldsymbol{\sigma}_@}$ for some values x_1, x_2 of X and y of Y.*

If X and Y are binary variables, then we can equivalently say that X is a probabilistic *total* cause of Y (in situation $\mathbf{u}$) if and only if,

$$\text{CE}(X = 1, Y = 1) \equiv P(Y(\mathbf{u}) = 1)_{X=1} - P(Y(\mathbf{u}) = 1)_{X=0} \neq 0. \tag{4.22}$$

[8]Proof: First, note that $Y(\mathbf{u}) = Y_{X=x}(\mathbf{u})$, where x is the *actual* value of X (i.e., the one determined by $\mathbf{U} = \mathbf{u}$). This is because the intervention on X alone doesn't change the *way* that any other variable depends on X, and we're considering an intervention that doesn't change the value of X; so the values of all other variables must be invariant also. Thus, 4.21 trivially entails Criterion 1. Now consider the other direction. Suppose Criterion 1 is satisfied. Then either x is distinct from both x_1 and x_2 or it isn't. If it isn't, then 4.21 is satisfied. If it is, then either $Y_{X=x_1}(\mathbf{u}) = Y_{X=x}(\mathbf{u})$ or the equality doesn't hold. If it holds, then $Y_{X=x_2}(\mathbf{u}) \neq Y_{X=x}(\mathbf{u})$, and 4.21 is statisfied with $x' = x_2$. If it doesn't, then 4.21 is satisfied with $x' = x_1$. QED.

Analogously, we can also say that X is a probabilistic *direct* cause of Y if and only if,

$$\mathrm{dCE}(X=1,Y=1) \equiv P(Y(\mathbf{u})=1)_{X=1,\mathbf{\Sigma}_{\mathrm{path}}=\boldsymbol{\sigma}_{@}} - P(Y(\mathbf{u})=1)_{X=0,\mathbf{\Sigma}_{\mathrm{path}}=\boldsymbol{\sigma}_{@}} \neq 0. \tag{4.23}$$

Note here that the probability distribution P over the endogenous variables remains non-extremal even after fixing definite values for all exogenous variables, $\mathbf{U}=\mathbf{u}$. This is a novel feature of the indeterministic generalization of structural causal models, and is not possible within the unmodified SCM formalism.

Unlike in the deterministic case, the alternative formalization that X is a probabilistic total cause of Y if and only if $P(Y_{X=x'}(\mathbf{u})=y) \neq P(Y(\mathbf{u})=y)$ is not equivalent to Intervention Criterion 3 (IC 3). This is because the first step of the proof of footnote 8, which establishes that $Y_{X=x}(\mathbf{u})=Y(\mathbf{u})$ for the actual value x of X, does not survive the transition to the indeterministic setting. In the indeterministic setting, $\mathbf{u}$ determines only the *chance* distribution over X rather than X's value, so intervening to fix the *value* of X using $\mathrm{do}(X=x)$ will generically have non-trivial consequences on the chance distributions of downstream variables. We therefore need to choose. I favour the criterion IC 3 (and its analogue for direct causation, IC 4), because, as we will see shortly, it plays well with the Mentaculus, giving us a common "branch point" from which to evaluate the associated counterfactuals.

4.3 Causal Models within the Mentaculus

At a high level, the logic of the project is now clear. The framework of causal models (CMs) establishes, via the intervention criterion, a tight connection between causal relationships and counterfactual claims. The Mentaculus, in turn, offers a reductive analysis of counterfactuals in terms of the probability $\Pr_{\mathcal{M}}(B(t')|A(t),M(t^*))$, where t^* is the branch-point time for $A(t)$. Thus, we can hope to use the Mentaculus both to recover various elements of the causal models framework (e.g., the directedness and acyclicity of the causal graph, and the rule of product decomposition) and also, beyond that, to justify the pervasive assumption of a uniform time-orientation for all correct causal graphs. The latter argument, presented in section 5.2, constitutes our fullest answer to the question of why causes always precede their effects in time.

By way of preliminaries, I will make a number of basic—but essential—remarks. First, recall that in section 3.2.2 we defined a macro*property* for a dynamical system to be a union of its macro*states*. We will continue to employ this terminology in application to the entire universe, but now we add that a macro-property predicated

of the universe *at a particular time* forms a macro-*proposition*. For instance, the macroproposition $A(t)$ says that the universe has macroproperty A at time t.

Second, it may have been noticed that discussion of the Mentaculus in Chapter 3 was couched in terms of macro-propositions, whereas discussion of causal models in this chapter has been couched in terms of random variables and their values. For convenience, we stipulate that a time placement is built into all our random variables, albeit usually suppressed in the notation. Thus, for instance, the statement "$Z = z$" might tell us the value z of someone's blood pressure *at a particular time*, the time of Z. Given this convention, and assuming that all the variables in question are macroscopic, the descriptions of the world in terms of the values of random variables and in terms of macropropositions are interchangeable.

Quick proof: A random (macro-)variable taking on a particular value (e.g.., $Y = y$) constitutes a particular macroscopic fact about the universe at some time. Any macroscopic fact about the universe at t supervenes on the universe's macrostate at t, though it may contain considerably less information and thus be compatible with a host of alternative universal macrostates-at-t. The union of all the compatible macrostates-at-t forms a macroproperty for the universe at t, and thus, in our terms, a macroproposition, $A(t)$. Conversely, any macro-proposition can be cast in terms of a random (macro-)variable taking on a particular value. For instance, for any macro-proposition $A(t)$ we can form the set $\{A(t), \neg A(t)\}$; we can then define the binary variable X with values in $\{0, 1\}$, where $X = 1$ indicates the truth of the proposition $A(t)$ and $X = 0$ indicates its falsity (i.e., indicates the truth of $\neg A(t)$). Often enough, however, the proposition $A(t)$ will have an intrinsic structure that allows for a more interesting correspondence with the random variable description.

Finally, CMs and the Mentaculus each give their own analysis of counterfactuals. Insofar as these respective analyses latch onto the same thing, and using the above correspondence between macro-propositions and the values of macro-variables, we can hope to devise a "translation scheme" between the formal language of CMs and the language of the Mentaculus. This is what is attempted in this section.

Let us begin by revisiting an earlier discussion. In section 3.2.1, we defined a probabilistic semantics for counterfactuals grounded in the Mentaculus. According to Counterfactuals 2, $A(t)\square\!\!\rightarrow \Pr(B(t')) = x$ is true just in case $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = x$, where t^* is the branch-point time for $A(t)$. It may have occurred to the reader that we passed over a natural alternative suggestion, namely a schema for *probabilities*

of definite counterfactuals rather than *counterfactuals with probabilistic consequents.* Consider therefore the following two versions of a probabilistic semantics:

$$\Pr(A(t) \boxdot\!\!\rightarrow B(t')) = x \text{ iff } \Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = x. \tag{4.24}$$

$$A(t) \boxdot\!\!\rightarrow \Pr(B(t')) = x \text{ iff } \Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = x. \tag{4.25}$$

To focus our analysis, let us take $A(t)$ to be the proposition that I toss a certain (fair) coin at t, and $B(t')$ to be the proposition that the coin displays heads at t' (which must, of course, be chosen to lie sufficiently after the time t of the tossing such that the coin will have landed and settled down). The branch-point time t^* is the latest time at which I might have decided not to toss the coin, and $M(t^*)$ is therefore the actual macrostate at this time. $M(t^*)$ is by definition compatible both with tossing and not tossing the coin at t, so it is extremely plausible that it is also compatible with a variety of different *ways* of tossing the coin. We can therefore infer that $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*)) = 1/2$.[9]

Now, what should we say about this case? According to the semantic schema 4.24, we should say that the counterfactual "if I had tossed the coin at t, it would have landed heads" is true with probability 1/2. By contrast, according to the semantic schema 4.25, we should say instead that a different, probabilistic counterfactual ("if I had tossed the coin at t, the probability of its landing heads would have been 1/2") is true. Which of these, if either, should be adopted in the indeterministic setting?

Pre-theoretical intuitions on this question may vary. Fortunately, however, there are good theoretical reasons to favour the semantic schema 4.25 over the schema 4.24. The semantic schemas (4.24) and (4.25) differ in that the former continues to work with definite counterfactuals (which have some probability of being realised in a given case), while the latter eschews definite counterfactuals entirely in favour of counterfactuals with probabilistic consequents. The semantics (4.24) therefore represents a probabilistic view onto an event space composed of definite counterfactuals, which is reminiscent of how we think about the causal effect in the *structural* (SCM) framework (see the end of sub-section 4.1.3). It is therefore natural to propose the following formal interpretation:

[9]The sort of argument to be adduced for this is along the lines of that in Strevens (2011). Roughly, we can expect the region of the phase space corresponding to slightly different initial conditions for the tossed coin (i.e., slightly different height from the ground, vertical momentum, and angular momentum) to be rapidly alternating in the eventual outcome of the coin toss (i.e., heads or tails) that is determined by the Hamiltonian dynamics. Thus, if the Mentaculus probability distribution conditional on $M(t^*)$ and $A(t)$ is somewhat randomly spread across the relevant part of the phase space, it will overlap with the portion of the phase space corresponding to "heads"-destined initial conditions to approximately the degree 1/2.

SCM Semantics 1 (Epistemic)**.** $\Pr(A(t) \boxdot\!\!\rightarrow B(t')) = x$ *iff* $P_{SCM}((Y = y)_{X=x'}|\mathbf{\Sigma} = \boldsymbol{\sigma}) = x$*, for a subset* $\mathbf{\Sigma} \subseteq \boldsymbol{V}$ *of the endogenous variables ("the evidence").*

I have called this interpretation "epistemic". Any given counterfactual pertains to a particular, token sequence of events (i.e., to what I call a particular *case*). In any particular case, the exogenous variables must all take on some particular values $\mathbf{U} = \mathbf{u}$, and this will engender determinate answers to all causal queries. We obtain non-extremal probabilities for $P_{SCM}((Y = y)_{X=x'}|\mathbf{\Sigma} = \boldsymbol{\sigma})$ only when treating the determinate values of the exogenous variables as *uncertain*, and consequently treating the conditioned-upon endogenous variables $\boldsymbol{\sigma}$ as *evidence* for these determinate-yet-unknown values of the $\mathbf{U}$-variables. Thus, an intermediate value for $P_{SCM}((Y = y)_{X=x'}|\mathbf{\Sigma} = \boldsymbol{\sigma})$ can come about only as a result of our uncertainty over the values of the $\mathbf{U}$-variables in the particular case, and the probability P_{SCM} can only represent an *epistemic probability*, quantifying the degree to which the evidence $(\mathbf{\Sigma} = \boldsymbol{\sigma})$ evidentially supports the definite counterfactual in question.[10]

By contrast, the Mentaculus probability $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*))$ involves conditionalizing on the full set of macroscopic facts at t^*, without restriction. Moreover, its value depends solely on the macro-propositions that are fed into the Mentaculus probability function, quite independently of what we know or can observe. The Mentaculus probability $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*))$ is therefore of the wrong *type* to correspond to $P_{SCM}(B(t')_{A(t)}|\mathbf{\Sigma} = \boldsymbol{\sigma})$; the former is a chance whereas the latter is an epistemic probability. Given the SCM interpretation of $\Pr(A(t) \boxdot\!\!\rightarrow B)$ (i.e., SCM Semantics 1), we must reject the semantic schema 4.24.

Cases of genuinely *probabilistic causation*, where the probabilities do not result merely from partial observability or ignorance, aren't well modelled in the framework of SCMs. To allow for particular, token counterfactuals of the form "if I had tossed the coin at t, it would have landed heads with probability 1/2" we require a different formalism, in which the analogue of the deterministic structural equations are themselves probabilistic relations. This is what is accomplished by the formalism of indeterministic causal models (ICMs), as outlined in section 4.1.4. With the ICM formalism in hand, the following causal-models interpretation suggests itself:

[10] In principle we could take $\mathbf{\Sigma} = \mathbf{V}$, the full set of endogenous variables. But note that, subject to the assumption of invertible structural functions, fixing values for all endogenous variables will also fix values for all the exogenous variables, and so once again only extremal probabilities can be obtained for $P_{SCM}((Y = y)_{X=x'}|\mathbf{V} = \mathbf{v})$. Thus, subject to this assumption, *any* proposal for an interpretation of $\Pr(A(t) \boxdot\!\!\rightarrow B(t'))$ in the SCM framework which is to obtain non-extremal probabilities requires some choice of a *strict* subset of the endogenous variables to condition upon. Since all the endogenous variables *in fact* take on definite values in any given case, this choice of subset can only be due to a lack of knowledge of some of these variables.

SCM Semantics 2 (Chance)**.** $A(t) \boxdot\!\!\rightarrow \Pr(B(t')) = p$ *iff* $P_{ICM}(Y(\boldsymbol{u}) = y)_{X=x'} = p$*, where* P_{ICM} *are the probabilities in an indeterministic causal model after fixing the exogenous variables to their actual values.*

Under this interpretation, it *would* be reasonable to demand, as a condition of adequacy on the indeterministic causal model, that $P_{ICM}(B(t'))_{A(t)}$ is equal to the Mentaculus probability $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*))$. While the former attempts to compute the *chance* of $B(t')$ under the counterfactual supposition of $A(t)$, given the particularities of the background circumstances ($\mathbf{U} = \mathbf{u}$), the latter is supposed to define what this chance is. Thus, the version (4.25) of the indeterministic semantics can be upheld, and this provides a retrospective justification for the choice of this semantic schema in section 3.2.1.

Putting together SCM Semantics 2 with the Mentaculus account of counterfactuals (schema 4.25), we obtain Mentaculus-based truth conditions for formal statements expressed in the language of causal models, which can act as constraints on the *interventional* side of causal model selection, over and above the *statistical* constraints imposed by the rule of product decomposition. In combination with the probabilistic intervention criteria for total and direct causal influence (IC 3, 4), we in fact obtain a method for *extracting* a full causal graph from the Mentaculus probabilities alone. This is illustrated in detail in the next chapter.

5

Explaining the Structure of Causality

The previous chapter developed the two ingredients necessary to relate the formal structure of causal models (and their associated causal graphs) to the time-asymmetric system of the Mentaculus: (1) a criterion for causal influence in terms of intervention counterfactuals (section 4.2), and (2) a set of truth conditions for those counterfactuals grounded in the Mentaculus (section 4.3). With these ingredients in place, we are now well-positioned to explain how and why the Mentaculus gives rise to systems of causal relationships with the formal structure of an (indeterministic) causal model, as well as to answer the question with which we began: what justifies the presumption of a time orientation for all correct causal graphs?

The chapter proceeds as follows. Section 5.1 demonstrates, with careful study of a simple example, how the ingredients developed in the previous chapter allow us to *extract* a causal graph from the Mentaculus. Section 5.2 then explains why this causal graph must be oriented with respect to the direction of time. Section 5.3 explains two further properties of the causal graph, commonly presupposed in the causal models formalism: its directedness and its acyclicity. Finally, section 5.4 offers a derivation of the contentious Causal Markov Condition.

5.1 Recovering the Causal Graph

For concreteness of analysis, let us return to our running example of Simpson's paradox. For convenience, we will stick to the notation using macropropositions rather than the notation using random variables. Thus, let $E(t)$ be the proposition that the subject (S) has elevated blood pressure at t; let $D(t')$ be the proposition

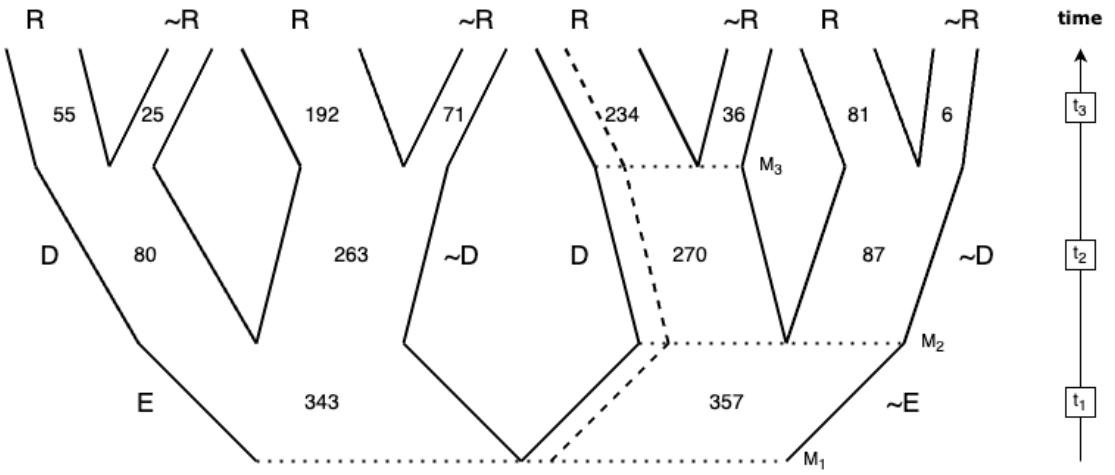

Figure 5.1: Schematic depiction of the relevant part of the Mentaculus "tree".

that S takes the drug at t'; and let $R(t'')$ be the proposition that S is recovered from their condition at t''. Suppose that in some particular case in the actual world, we have a situation where:

- $\neg E(t_1)$: At t_1, S's blood pressure is low (i.e., not elevated).
- $D(t_2)$: At t_2, S takes the drug.
- $R(t_3)$: At t_3, S is recovered from their health condition.

And let us suppose the time order $t_1 < t_2 < t_3$ for the events. From this time order, we assume that we can already infer that the correct causal structure is the one shown in figure 4.5 rather than the one shown in figure 4.1. Given that $t_1 < t_2$, we assume that there cannot be a causal influence of S's drug-taking on their earlier blood pressure. So the task we now have to take on is the following: can we *derive* the causal structure of figure 4.5 for this case, using only the intervention criterion and the Mentaculus semantics for counterfactuals? Once we have done this, we will be in a better position to see how and why the intervention criterion agrees with the verdict of time order on the question of the causal relationship between drug-taking and blood pressure.

Figure 5.1 depicts the situation according to the Mentaculus. The dashed line represents the world's *actual* micro-trajectory, and the dotted lines represent the macrostates at relevant branch points. The numbers in the branches are taken from table 1.1. For simplicity, we assume that the statistics manifested in the whole population ($N = 700$) provide a guide to the *chances* operative in each case, via the law of large numbers. Thus, while strictly speaking figure 5.1 only depicts

the situation for a *single individual* in the population (characterized by a full set of definite values for the exogenous variables $\mathbf{U} = \mathbf{u}$), we conveniently take the numbers in table 1.1 as estimates for the Mentaculus conditional probability of a sub-branch conditional on its parent branch *in that individual's particular case*. It is clear, then, that the only sort of causal model we can hope to recover from this will be of the indeterministic (ICM) variety.

To get our bearings, let us start with one slow and careful calculation of a causal effect from figure 5.1. I will then proceed more systematically, but more quickly, through the handful of other causal-effect calculations needed to recover the full causal graph of figure 4.5. Consider the *total causal effect* of S's low blood pressure on their recovery. According to Intervention Criterion 3, we need to compare $P(R(t_3))_{\neg E(t_1)}$ with $P(R(t_3))_{E(t_1)}$. These counterfactuals can be evaluated using the Mentaculus truth conditions for probabilistic counterfactuals (Counterfactuals 2). The first step, as always, is to identify the branch-point time. The branching-off of the counterfactual antecedent $E(t_1))$ from the actual antecedent $\neg E(t_1))$ occurs at a time shortly before t_1, when the macrostate is M_1. Thus, we compute the counterfactuals as follows,

$$P(R(t_3))_{E(t_1)} = \Pr_{\mathcal{M}}(R(t_3)|E(t_1), M_1) = 247/343 \tag{5.1}$$

$$P(R(t_3))_{\neg E(t_1)} = \Pr_{\mathcal{M}}(R(t_3)|\neg E(t_1), M_1) = 315/357. \tag{5.2}$$

The causal effect of lowered blood pressure on recovery is then given by,

$$\mathrm{CE}(\neg E(t_1), R(t_3)) = P(R(t_3))_{\neg E(t_1)} - P(R(t_3))_{E(t_1)} = 315/357 - 247/343 \approx 0.16. \tag{5.3}$$

We conclude that S's lowered blood pressure has a non-zero (and positive) *total* causal effect on their recovery.

To recover the full causal graph, we need to consider the full network of possible causal relations between all three macropropositions. In other words, for each pair of macropropositions in the set $\{\neg E(t_1), D(t_2), R(t_3)\}$, we need to apply the intervention criterion (IC 3) in both directions. Again, the intervention criterion is formulated in terms of counterfactuals for which the Mentaculus supplies truth conditions. The required Mentaculus probabilities are displayed below.

$$\begin{array}{lll}
(\neg E \to D) & \Pr(D|\neg E, M_1) = \dfrac{270}{357}, & \Pr(D|E, M_1) = \dfrac{80}{343} \\
(D \to \neg E) & \Pr(\neg E|D, M_2) = 1, & \Pr(\neg E|\neg D, M_2) = 1 \\
(\neg E \to R) & \Pr(R|\neg E, M_1) = \dfrac{315}{357}, & \Pr(R|E, M_1) = \dfrac{247}{343} \\
(R \to \neg E) & \Pr(\neg E|R, M_3) = 1, & \Pr(\neg E|\neg R, M_3) = 1 \\
(D \to R) & \Pr(R|D, M_2) = \dfrac{234}{270}, & \Pr(R|\neg D, M_2) = \dfrac{81}{87} \\
(R \to D) & \Pr(D|R, M_3) = 1, & \Pr(D|\neg R, M_3) = 1
\end{array} \tag{5.4}$$

(Here I have dropped the explicit reference to time in the notation to reduce clutter, though each macroproposition should of course still be interpreted as occurring at its respective time, as stipulated above.)

Using these probabilities to compute the causal effects, we obtain:

$$\begin{array}{lll}
\mathrm{CE}(\neg E, D) = 270/357 - 80/343 & \approx 0.52 \\
\mathrm{CE}(D, \neg E) = 1 - 1 & = 0 \\
\mathrm{CE}(\neg E, R) = 315/357 - 247/343 & \approx 0.16 \\
\mathrm{CE}(R, \neg E) = 1 - 1 & = 0 \\
\mathrm{CE}(D, R) = 234/270 - 81/87 & \approx -0.064 \\
\mathrm{CE}(R, D) = 1 - 1 & = 0.
\end{array} \tag{5.5}$$

Note that there is a causal path from $\neg E$ to R running over D, since both $\mathrm{CE}(\neg E, D)$ and $\mathrm{CE}(D, R)$ are non-zero; the *direct* causal effect $\mathrm{dCE}(\neg E, R)$ may therefore differ from the *total* causal effect $\mathrm{CE}(\neg E, R)$ just computed, and we need to investigate this further before drawing up the causal graph.

The direct causal effect $\mathrm{dCE}(\neg E, R)$ can be computed using Intervention Criterion 4, by comparing $P(R)_{E,D}$ with $P(R)_{\neg E,D}$. Using the Mentaculus semantics, this gives us,

$$\mathrm{dCE}(\neg E, R) = \Pr(R|\neg E, D, M_1) - \Pr(R|E, D, M_1) = 234/270 - 55/80 \approx 0.18. \tag{5.6}$$

Comparison with equation 5.3 reveals that the direct causal effect differs only slightly from the total causal effect: it is also positive, but with a slightly larger magnitude. This result can also be inferred from equations 5.5: since D causally lowers the probability of R, but $\neg E$ causally raises the probability of D, the indirect causal effect of $\neg E$ on R (i.e., that part of it running through D) must be *negative*. Since the total causal effect represents the combined causal effect along the direct and indirect routes, the *direct* causal effect must be larger to compensate.

Equations 5.5 do not reveal any other indirect causal paths, so the remaining direct effects must be equal in value to their corresponding total effects, and we are now in a position to draw up the causal graph. It is shown in figure 5.2.

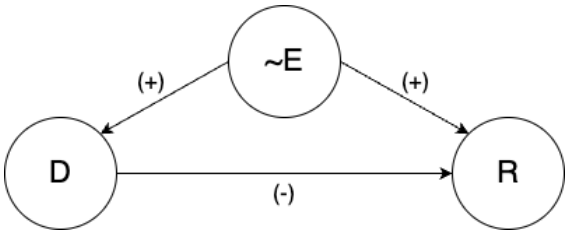

Figure 5.2: Causal graph as recovered from figure 5.1. This graph matches that of figure 4.5 up to the omission of the exogenous variables, which are fixed constants of the particular case, and the replacement of the endogenous variables with their actual values.

A few comments are in order on the interpretation of this graph. Note first that, unlike the causal graphs in earlier chapters, the nodes in this graph do not stand in for the variables simpliciter, but rather for their *actual values*. For example, the Z node of figure 4.5 becomes the "$\neg E$" (or $Z = 0$) node in figure 5.2. Consequently, the edges obtain a polarity (+ or −), indicating whether the actual value of the parent node causally promotes or causally inhibits the actual value of the daughter node. Drawing the graph in terms of the actual values of the variables also signifies that the causal effects have been calculated *with respect to* those values. This can make a difference to their exact magnitudes, though importantly it has no impact on the structure of the graph.[1] Causal effects computed from the Mentaculus always pertain to an "actual" graph such as figure 5.2 rather than a general graph such as figure 4.5. (There is no philosophical consensus on how precisely to define so-called "actual causation", but a range of related discussions can be found in Woodward (2004, pp. 74–86), Halpern (2016, pp. 23–27), and Glymour and Wimberly (2007).)

Second, this graph omits the exogenous variables. This is because we have taken the exogenous variables as fixed constants of the particular case under study—they contribute towards the values of the transition chances (i.e., the Mentaculus

[1]For example, a calculation of the causal effect of D on R based on the general graph in figure 4.5 would yield, using the adjustment formula (equation 4.10),

$$\begin{aligned}\mathrm{CE}(D,R) &= P(R|\mathrm{do}(D)) - P(R|\mathrm{do}(\neg D)) \\ &= P(E)(P(R|D,E) - P(R|\neg D,E)) + P(\neg E)(P(R|D,\neg E) - P(R|\neg D,\neg E)) \\ &= \frac{343}{700}\left(\frac{55}{80} - \frac{192}{263}\right) + \frac{357}{700}\left(\frac{234}{270} - \frac{81}{87}\right) \approx -0.0537.\end{aligned}$$

By contrast, when calculating the causal effect from the values-specific graph (figure 5.2), we send $P(E) \to 0$, $P(\neg E) \to 1$, to obtain $\mathrm{CE}(D,R) = \frac{234}{270} - \frac{81}{87} \approx -0.064$, thereby extracting just that *part* of the causal effect that is operative in the actual world, in which $\neg E$. What is crucial, however, is that in both cases the Z (or $\neg E$) node is treated as a parent, and the causal effect is calculated accordingly, by conditioning on it. This is why these differences in the exact values of the causal effects computed from figure 4.5 versus 5.2 make no difference to the graph structure.

conditional probabilities), but they are not themselves included in the model we have extracted. This is to be expected, since figure 5.1 assumes access to the full facts of the particular case, so we have no uncertainty $P(\mathbf{u})$ over the background conditions to consider and all probabilities in the model come from the chancy transitions between variables. We were able to get rid of the epistemic component $P(\mathbf{u})$ with our stipulation that the statistics of table 1.1 provide estimates for the *chances* operative in each particular case. Of course, in practice, these statistics would be better interpreted as providing estimates of the probabilities that result from the "mixing" of the chance transition functions (which we identify with the Mentaculus conditional probabilities) with a probability distribution over the **U**-variables (representing our uncertainty about which case we're in); see section 4.1.4 for an elaboration.

With these comments in mind, comparison with figure 4.5 shows that we have recovered the expected causal graph for the situation. The treatment of this example demonstrates how one may go from assumptions about the time order of events, via the Mentaculus and the intervention criterion, to a full causal graph of the situation.

It is instructive to briefly compare these results with those we would have obtained had we assumed a different time order for the events. Thus suppose instead that we have a situation where,

- $D(t_1)$: At t_1, S takes the drug.
- $\neg E(t_2)$: At t_2, S's blood pressure is low (i.e., not elevated).
- $R(t_3)$: At t_3, S is recovered from their health condition.

The associated part of the Mentaculus tree is depicted in figure 5.3.

Following the same procedure as above, we can compute the total causal effects, in both directions, between each pair of variables. These come out as:

$$\begin{aligned}
\mathrm{CE}(\neg E, D) &= 1 - 1 && = 0 \\
\mathrm{CE}(D, \neg E) &= 270/350 - 87/350 && \approx 0.52 \\
\mathrm{CE}(\neg E, R) &= 234/270 - 55/80 && \approx 0.18 \\
\mathrm{CE}(R, \neg E) &= 1 - 1 && = 0 \\
\mathrm{CE}(D, R) &= 289/350 - 273/350 && \approx 0.046 \\
\mathrm{CE}(R, D) &= 1 - 1 && = 0.
\end{aligned} \tag{5.7}$$

Once again, we must check for possible indirect causal paths. Here we note that both $\mathrm{CE}(D, \neg E)$ and $\mathrm{CE}(\neg E, R)$ are non-zero, and so there is an indirect path from D to R, running over $\neg E$. Thus, the direct causal effect $\mathrm{dCE}(D, R)$ may differ from $\mathrm{CE}(D, R)$, and we must investigate this before drawing up the causal graph.

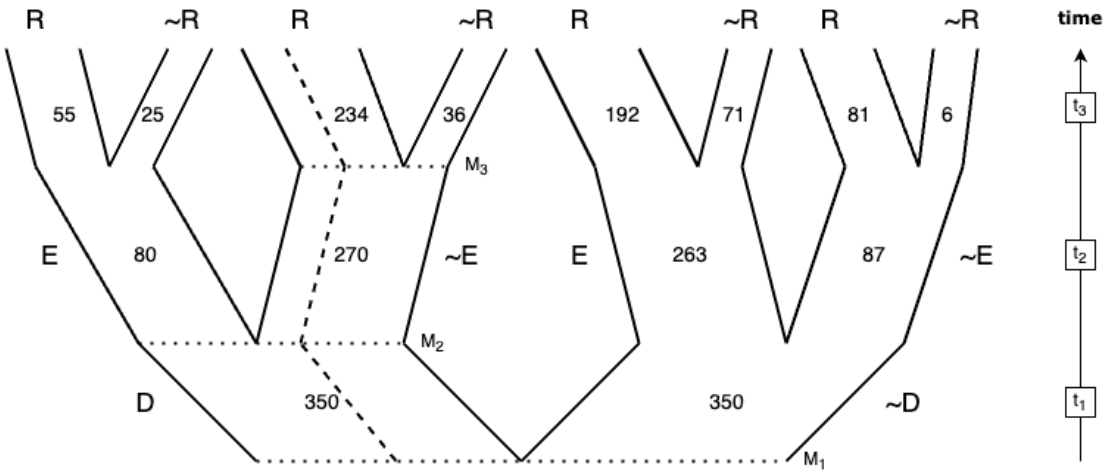

Figure 5.3: Schematic depiction of the Mentaculus "tree" with reversed time-order for $\neg E$ and D.

For the direct causal effect of D on R, we compute,

$$\begin{aligned} \mathrm{dCE}(D,R) &= P(R)_{D,\neg E} - P(R)_{\neg D,\neg E} && (5.8)\\ &= \Pr(R|D,\neg E,M_1) - \Pr(R|\neg D,\neg E,M_1) && (5.9)\\ &= 234/270 - 81/87 \approx -0.064. && (5.10) \end{aligned}$$

Here we have a more interesting phenomenon: the *reversal of sign* of the direct causal effect $\mathrm{dCE}(D,R)$ from the total causal effect $\mathrm{CE}(D,R)$. As in the previous case, the *direct* causal effect of the drug on recovery is negative (and indeed has the same value). However, whereas in the previous case the direct causal effect was found to be *equal* to the total causal effect, in this case we find that the total causal effect is net positive, due to the indirect path running over $\neg E$. Again, the story naturally emerges from equations 5.7: $\mathrm{CE}(D,\neg E) > 0$, so the drug causally promotes lowered blood pressure; additionally, $\mathrm{CE}(\neg E,R) > 0$, so lowered blood pressure causally promotes recovery. In combination, the drug therefore aids recovery by lowering blood pressure. The total beneficial effect of the drug is, however, reduced somewhat by a negative direct causal effect—some mildly toxic side effect.

The causal graph for this case therefore takes the form of figure 5.4. Again, comparison with figure 4.1 shows that we have recovered the expected causal graph, up to the omission of the exogenous variables and the replacement of endogenous variables with their actual values. We have therefore rediscovered through the Mentaculus what we intuitively reasoned must be the case in Chapter 1, and moreover we have found that even the quantitative results of the Mentaculus analysis agree, after accounting for the distinction between actual and general causation, with the results of the causal-models analysis given in Chapter 4.

In sum, the detailed treatment of this specific example exemplifies a procedure for how one may recover a causal graph from the Mentaculus, and gives us an insight into how the time order of events helps to determine the form of this graph. With this example in mind, we are now well-placed to seek a more general explanation for the time orientation of causal graphs.

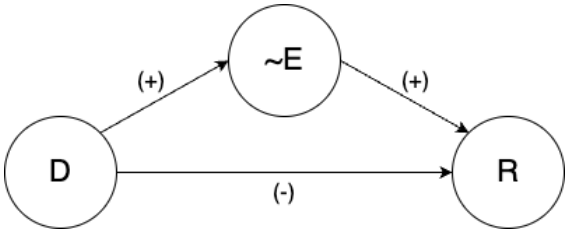

Figure 5.4: Causal graph as recovered from figure 5.3. This graph matches that of figure 4.1 up to the omission of the exogenous variables and the replacement of the endogenous variables with their actual values.

5.2 Time Orientation

In general terms, the Mentaculus semantics for counterfactuals, in conjunction with the intervention criterion, equates causal dependence between two macropropositions (or macrovariable-values) with their probabilistic dependence *conditional on the branch-point macrostate*. Note that if we were to omit the conditioning on the branch point, then for any two macropropositions $A(t)$, $B(t')$, we would have,

$$\begin{aligned}\Pr(B(t')|A(t)) \neq \Pr(B(t')) &\iff \Pr(A(t), B(t')) \neq \Pr(A(t))\Pr(B(t')) \\ &\iff \Pr(A(t)|B(t')) \neq \Pr(A(t)).\end{aligned}$$

Consequently, counterfactual dependence would necessarily be a symmetric relation, inherited from the symmetry of simple probabilistic dependence. The key to the asymmetry, according to the Mentaculus account, is the differential placement of the branch-point macrostate when computing the counterfactual dependence of $A(t)$ on $B(t')$ as compared with that of $B(t')$ on $A(t)$.

Return to the example of figure 5.1. When computing the *forwards* counterfactual dependence of S's drug taking on their blood pressure, $\Pr(D(t_2)|E(t_1), M(t^*))$, we find that the branch-point time t^* must be placed earlier than *both* the antecedent and the consequent. Due to the forwards-branching structure of the Mentaculus, t^* cannot be placed any time *later* than the time of the antecedent; once the tree has branched into $E(t_1)$ and $\neg E(t_1)$, those branches remain $E(t_1)$ and $\neg E(t_1)$ forevermore, and the *actual* macrostate $M(t^*)$ is thus incompatible with the (contrary-to-fact) counterfactual antecedent $E(t_1)$ for all $t^* > t_1$. So, in particular, since the consequent occurs later in time than the antecedent, the branch-point time cannot be placed later than the time of the consequent, either. The only time order consistent with the branching structure of the Mentaculus is therefore

$t^* < t_1 < t_2$. This is why we computed the relevant counterfactuals, in section 5.1, by conditioning on the macrostate M_1.

By contrast, when computing the reverse, *backwards* counterfactual dependence of S's blood pressure on their drug taking, $\Pr(E(t_1)|D(t_2), M(t^*))$ a different possibility presents itself. By stipulation of the case, $t_1 < t_2$, so *prima facie* there are three possibilities for the placement of t^*. Once again, the forwards-branching structure of the Mentaculus immediately rules out the possibility $t_2 < t^*$: once the world has "chosen" $D(t_2)$ at t_2, $M(t^*)$ is forevermore incompatible with $\neg D(t_2)$. That leaves the possibilities $t^* < t_1$ and $t_1 < t^* < t_2$. The former obtains only if there is no time after the blood pressure measurement at which S "could have" decided not to take the drug (i.e., in a way which comports with our expectations for thermodynamically lawful macroscopic behaviour). But this is implausible. Decisions originate in the depths of S's brain, so there is presumably nothing about the *macrostate* at t_1 which necessitates their taking the drug at t_2. Presumably, indeed, S could have changed their mind at the very last second before t_2. Most plausibly, then, it is $t_1 < t^* < t_2$ that is the correct placement. This is why we computed the relevant counterfactuals, in section 5.1, by conditioning on the macrostate M_2 rather than M_1.

At bottom, then, we can trace the whole time-orientation of counterfactual dependence, and hence of causation, to the following simple fact: for any arbitrary counterfactual $A(t) \boxdot\!\!\rightarrow \Pr(B(t')) = p$, evaluated using the Mentaculus probability $\Pr_{\mathcal{M}}(B(t')|A(t), M(t^*))$, the placement of t^* follows the schema,

$$\begin{cases} t^* < t < t' \text{ if } t < t' \\ t' < t^* < t \text{ if } t' < t. \end{cases} \tag{5.11}$$

This difference in the placement of the branch-point time has the following implications for the derived causal structure.

For the case of *forwards* counterfactuals ($t^* < t_1 < t_2$), the branch-point macrostate $M(t^*)$ is generically compatible (in the sense of non-negligible Mentaculus conditional probability) with all four possible combinations of antecedent and consequent: $E(t_1)\&D(t_2), E(t_1)\&\neg D(t_2), \neg E(t_1)\&D(t_2), \neg E(t_1)\&\neg D(t_2)$. We find that $D(t_2)$ is counterfactually *independent* of $E(t_1)$ only in the special case that the $E(t_1)$ and $\neg E(t_1)$ branches are *symmetric* with respect to the ratio of probability given to $D(t_2)$ and $\neg D(t_2)$; otherwise, $D(t_2)$ is counterfactually *dependent* on $E(t_1)$ (as indeed was the case in the example treated in section 5.1).

By contrast, for the case of *backwards* counterfactuals ($t_1 < t^* < t_2$), the branch point macrostate $M(t^*)$ is generically only compatible with *two* of the four

possible combinations of antecedent and consequent—namely, the two possibilities involving the *actual* macroproperty at t_1. (In our example, these are $\neg E(t_1)\&D(t_2)$ and $\neg E(t_1)\&\neg D(t_2)$.) Thus, given that in the actual world $\neg E(t_1)$ in true, then irrespective of whether it were the case that $D(t_2)$ or $\neg D(t_2)$, the probability of $E(t_1)$ would've been zero and the probability of $\neg E(t_1)$ would've been one. The intervention criterion (IC 3) is therefore not satisfied, and we do not find a causal influence of the later event (drug-taking) on the earlier event (blood pressure).

This sort of argument against retrocausal influence generalizes quite far. The crucial step in the argument is the placement of the branch-point time t^* between the time of the counterfactual consequent and the time of the antecedent, $t_1 < t^* < t_2$. Once this is secured, the branching structure of the Mentaculus takes care of the rest, driving down the Mentaculus conditional probability of contrary-to-fact macropropositions at t_1 to a negligible value, and driving up the Mentaculus conditional probability of in-accordance-with-fact macropropositions at t_1 to a near certainty.[2]

Under what conditions, then, is this particular placement for t^* secured? The question comes down to the balance of two factors: (1) the size of the temporal interval between t_1 and t_2, and (2) the amount of "backtracking" required to lawfully transition from $M(t^*)$ to $A(t_2)$. Holding (2) fixed, there will be a smallest interval $t_2 - t_1$ beneath which retro-active intervention counterfactuals of the above form might start to come out true. Holding (1) fixed, there will be a maximum level of departure from the actual facts (as specified in the counterfactual antecedent $A(t_2)$) beyond which we can expect that the minimum required backtracking to $M(t^*)$ (i.e., the short stretch of the past macro-history that must be re-written to realise $A(t)$) takes us back even earlier than t_1.

Note that both factors are a function of the sort of causal relationships being probed, and hence of the sort of causal explanations being sought. While there would seem to be no general constraints on (1), the size of the temporal interval between the intervention and the desired effect (other than, perhaps, the human capacity to discriminate time order, unaided or otherwise), there is, intuitively, a salient constraint on (2). Causal models are not usually geared towards investigating the counterfactual consequences of dramatic, wholescale transformations to the macrostate of the universe. Rather, they are geared towards investigating the consequences of the sorts of modest, localized manipulations of the world around us that we can realistically effect as limited human agents. The sorts of variables

[2]Again, "negligible value" means probability zero in the thermodynamic limit, and "near certainty" means probability one in this limit.

figuring in a typical causal model are therefore almost invariably highly *localised*, occurring at a particular place in space rather than spread out across large regions.

Interventions on these localized variables constitute only a very limited departure from the totality of actual macroscopic facts at the time of the intervention. The degree of backtracking along the Mentaculus tree needed to account for such interventions is thus expected to be correspondingly limited. The present analysis, together with the seemingly universal alignment of causal relationships with the arrow of time, therefore tells us something important about human causal explanations, albeit something we already knew: that humans are especially interested in modeling the effects of what amount to only fairly modest, tightly localised manipulations of the world's macro-condition.

5.3 Directedness and Acyclicity

Establishing a time orientation for all causal relationships—or at least for all those causal relationships that might obtain amongst the localized variables of practical interest—has some important immediate consequences.

Consider a causal model of a set of variables all located at different times. Imposing a time-order criterion entails that the causal model satisfies two basic properties.

First, causal influence between the variables must be an asymmetric relation. Suppose for contradiction that X causes Y *and* Y causes X. If X causes Y, then by the time order criterion, X occurs before Y. Similarly, if Y causes X, then by the time order criterion X occurs after Y. But X cannot occur both strictly before and strictly after Y. So either X doesn't cause Y or Y doesn't cause X. The asymmetry of causal influence shows up as the *directedness* of all the edges in the causal graph.

Second, there can be no causal loops. Suppose someone were to propose a causal model of Simpson's paradox of the form depicted in figure 5.5, in which drug-taking causes recovery, recovery causes lowered blood pressure, and lowered blood pressure in turn causes drug-taking. As applied to any single case, this model is incompatible with the requirement that the directions of all edges are oriented uniformly with respect to time. Since the structure of time is (at least locally) acyclic, the causal graph cannot

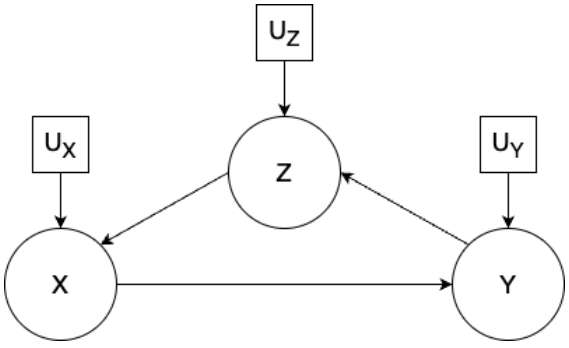

Figure 5.5: A cyclic causal graph.

be cyclic while also maintaining a consistent orientation with respect to the arrow of time. Thus, in figure 5.5, one of the edges XY, YZ, or ZX must point backwards in time, in violation of the time order criterion. The property of causal graphs that precludes loops like in figure 5.5 is called the *acyclicity* of the causal graph.[3]

5.4 The Causal Markov Condition

As illustrated by the example of Simpson's paradox, causal models go beyond statistical data insofar as they encode information about the effects of hypothetical interventions, and multiple distinct causal models may be compatible with the same dataset. However, as described in section 4.1, a statistical dataset does place certain *constraints* on the admissible causal models of a variable set $\mathbf{V}$. One such constraint is the so-called "rule of product decomposition", also called *factorizability* (Pearl et al. 2016, p. 29). This rule asserts that, given a causal graph of a variable set $\mathbf{V}$, the joint probability distribution over the variables can be factorized as:

$$\Pr(X_1, X_2, ..., X_n) = \prod_i \Pr(X_i|\mathbf{PA}_i), \tag{5.12}$$

...where the product runs over all the variables and $\mathbf{PA}_i$ stands for the parents of X_i.

If the graph is acyclic, then we can always relabel the variables so that X_j is causally downstream of X_i only if $i < j$ (in other words, the indices are strictly increasing along every directed path). Then, using the definition of conditional probability, we can write:

$$\Pr(X_1, X_2, ..., X_n) = \Pr(X_1) \Pr(X_2|X_1)... \Pr(X_n|X_{n-1}, X_{n-2}, ..., X_2, X_1). \tag{5.13}$$

Since the variables have been relabelled such that the indices are strictly increasing along every directed path, each factor in equation 5.13 has the form $\Pr(X_i|\Sigma_i)$, where Σ_i is a set (not necessarily complete) of *non-descendants* of the variable X_i. What's more, although Σ_i need not contain *all* the non-descendants of X_i, it must contain all the *ancestors* of X_i (and hence, a forteriori, $\mathbf{PA}_i$). For since Σ_i contains all the variables with indices $j < i$, if there were an ancestor X_k missing from Σ_i, it could only have $k > i$, in violation of the requirement that X_i is causally downstream of X_k only if $k < i$. Equation 5.12 is thus derivable if we assume the following independence condition:

[3] I say that the time order criterion mandates that the causal graph be *locally* acyclic so as not to preclude the possibility that time is indeed cyclic at the level of the whole universe, as some models of cosmology allow. Note, however, that even violations of *global* acyclicity produce problems for the factorizability of the joint probability distribution over $\mathbf{V}$, discussed below.

$$\Pr(X_i|\mathbf{PA}_i, Y) = \Pr(X_i|\mathbf{PA}_i), \tag{5.14}$$

...where Y is any non-descendant of X_i (i.e., any variable that is not an effect of X_i). This independence condition is known as the *Causal Markov Condition*.[4]

In fact, it can be shown that in the case of acyclic graphs, equation 5.14 is not only sufficient but also necessary for factorizability (Hausman and Woodward 1999). The Causal Markov Condition (CM) therefore encapsulates the entire link between causality and statistics. CM is traceable back to Reichenbach's Common Cause Principle (H. Reichenbach and M. Reichenbach 1999, p. 163), and has been hotly debated in the literature ever since (see, for example, the objections raised against it by Forster (1988), Arntzenius (1992), and Cartwright (2002)). There is, consequently, great interest in understanding just to what degree and in what manner it can be justified. Here, too, the Mentaculus can provide insight.

Since equation 5.14 relates conditional probabilities, it is tempting to try to derive it directly from the Mentaculus formula for conditional probabilities,

$$\Pr_{\mathcal{M}}(B(t')|A(t)) \equiv \frac{\int P(B)L(t'-t)P(A)L(t)\rho_0 \, \mathrm{d}\mu}{\int P(A)L(t)\rho_0 \, \mathrm{d}\mu}. \tag{5.15}$$

The trouble is that the Mentaculus doesn't seem, in itself, to tell us anything general about patterns of conditional (in)dependence among macro-propositions. There is, however, a suggestive analogy between the Causal Markov Condition and the Markovian *forwards predictability* of the macroscopic dynamics, which the Mentaculus helps to underwrite (see section 3.2.2). Recall that in statistical mechanics, information about history is encoded in microscopic correlations in the present state, and that applying a Gibbs coarse-graining involves discarding this information. Recall also that, provided we are prepared to grant Wallace's *Simple Dynamical Conjecture* (SDC), applying a Gibbs coarse-graining to the Mentaculus distribution over the world's macrostate commutes with applying the exact distributional time-evolution (i.e., the Liouville dynamics) *forwards* in time. The Mentaculus therefore assures us that there is enough information contained within the *present* macrostate alone in order to make further details of the macro-history probabilistically irrelevant to the macro-future.

What the Mentaculus *does not* tell us is which particular macro-propositions about the present are relevant to the prediction of which particular propositions about the future. Prima facie, it might seem that the Causal Markov Condition tells us more about this. But appearances can be misleading.

[4]Spirtes et al. (1993); this formulation is taken from Hausman and Woodward (1999).

Consider the following (adapted) example due to Wesley Salmon, concerning the causal interactions among some billiard balls. Let C represent whether a cue ball collides ($C = 1$) or does not collide ($C = 0$) with two mutually adjacent billiard balls. Let $E1$ represent whether the first ball is pocketed, and $E2$ represent whether the second ball is pocketed. The causal graph shown in figure 5.6. Let us now stipulate that the configuration of the balls is such that the first ball will only be nudged into the pocket if the second is also nudged into its pocket. Thus, although C is the only parent of $E1$, and $E2$ is not a descendant of $E1$, we still find that $E1$ is probabilistically dependent on $E2$ conditional on C, in violation of the Causal Markov Condition (Salmon 1985, p. 168).

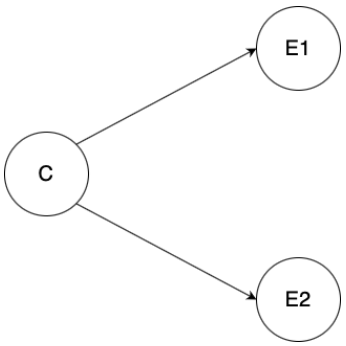

Figure 5.6: A violation of the Causal Markov Condition?

So, does the Causal Markov Condition fail? With respect to the variable set $\mathbf{V} = \{C, E1, E2\}$, it does indeed. But, as Hausman and Woodward (1999) hasten to add, it is still possible to satisfy the Causal Markov Condition with respect to a different parameterization of the situation. Suppose that, instead of C, we were to include C', the exact momentum of the cue ball upon impact, in the variable set. Then indeed conditionalization on C' (the new parent of $E1$) would screen off the correlation between $E1$ and $E2$, and the validity of the Causal Markov Condition would be restored.

Let us call a causal model of a given situation *correct* if it is in accordance with the intervention criterion for causal influence. Then we can put the lesson of the preceding considerations as follows: the true statement is not that CM must hold with respect to every correct causal model of the situation, involving any choice of variable set $\mathbf{V}$; the true statement is rather that there always exists a correct causal model of the situation with respect to *some* variable set $\widetilde{\mathbf{V}}$ for which CM is satisfied.[5] In light of this, it is hardly surprising that the Mentaculus gives us no reason to think that CM holds with respect to arbitrary choices of macro-propositions (i.e., propositions about the values of arbitrary choices of macro-variables). It couldn't do this, because CM doesn't hold for arbitrary variable sets! What the Mentaculus assures us of is precisely what is needed: that it is always possible to find, within

[5]As defined here, "correct" simply means the same as "in accordance with the intervention criterion". Of course, provided that there always exists a correct causal model for which CM is satisfied, we could consider incorporating CM into the definition of what it is for a causal model to be "correct". We would then say that only a causal model that satisfies *both* the intervention criterion *and* CM can be considered correct.

any given macrostate, a choice of macro-variables that act as screening-off common causes for the purposes of predicting the probabilities of different macro-futures. The Mentaculus thus helps to explain, not why CM must hold in *every* correct causal model (which it doesn't), but why there must always *exist* a correct causal model with respect to which CM holds.

As always, an example will help to bring the idea into focus. There has been much debate in the philosophical literature concerning the rational choice in a curious dilemma known as *Newcomb's problem* (Nozick 1969). The set-up is as follows.

> Newcomb's Problem (NP): A subject is standing alone in a room, faced with a choice. In front of her are two boxes. The first box, which is opaque, contains either \$1M or nothing; the second, which is transparent, visibly contains \$1000. The subject may either take just the first box (a strategy called "one-boxing") or take both boxes ("two-boxing"), and keep whatever money she finds inside. The catch is that a prediction has been made in advance about what she will do, and the opaque box has been filled just in case it was predicted that she will one-box. It is stipulated that the predictor is highly reliable, and that the subject knows this. What should she do?

According to one tradition in decision theory, what she should do depends on the causal structure of her situation.[6] But just what is this causal structure? Let $B \in \{\text{one-box}, \text{two-box}\}$ be a variable representing whether the subject one-boxes or two-boxes; let $P \in \{\text{one-box}, \text{two-box}\}$ be a variable representing the outcome of the prediction for the subject's action, and let $M \in \{\$0, \$1000, \$1\text{M}, \$1\text{M} + \$1000\}$ be the amount of money that the subject takes home. Then, a first suggestion for a possible causal model might look like figure 5.7. (N.B., I have chosen the vertical positioning of the nodes to correspond to their time order. M has no exogenous contribution because it is completely determined by B and P.)

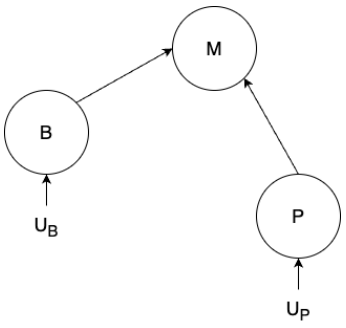

Figure 5.7: Newcomb's Problem, first suggestion.

[6] This tradition is of course *causal decision theory*. Causal decision theorists usually say that the subject should two-box. By the time the subject makes her choice, the opaque box has either already been filled or it hasn't; by taking both boxes, the subject takes away an additional \$1000 regardless of what was predicted. By contrast, *evidential* decision theorists usually say that the subject should one-box. Conditional on one-boxing, the probability that the opaque box is filled with \$1M is much higher (given the reliability of the predictor); the subject therefore maximises her expected winnings by taking only one box. For an entertaining survey of the debate between causal and evidential decision theories, see Lewis (1981). I will not here weigh in on this debate; my concern is instead to clarify just what causal structure may be assumed for this case.

Applying the Causal Markov Condition (CM), we immediately see that this suggestion will not do. CM says that for all distinct variables X and Y in the (endogenous) variable set $\mathbf{V}$, if X does not cause Y, then $\Pr(X|\mathbf{PA}_X, Y) = \Pr(X|\mathbf{PA}_X)$. (As usual, we take $\mathbf{PA}_X \subset \mathbf{V}$.) In this causal model, B does not cause P and $\mathbf{PA}_B = \varnothing$, so application of CM would lead us to conclude that $\Pr(B|P) = \Pr(B)$. By contrast, I stipulated the case such that the predictor is *reliable*, and a fortiori that their predictions are better than random chance, $\Pr(B|P) \neq \Pr(B)$. Thus, by the criterion CM, figure 5.7 cannot be the correct causal model.

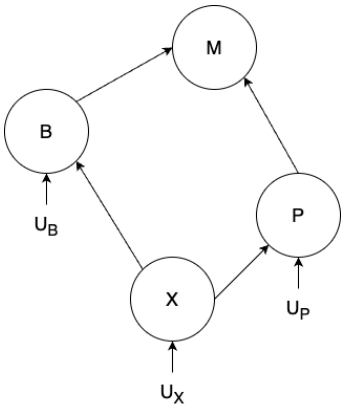

Figure 5.8: Newcomb's Problem with a "common cause" structure.

How the causal model be amended to satisfy CM? There are three options: either we introduce a causal arrow from P to B, or from B to P, or we enrich the variable set $\mathbf{V}$ with a "common cause" macro-variable, which we may denote X. In keeping with the spirit of the example, which is intended to drive a wedge between causal and evidential decision theories, we assume that the set-up has been described in whatever way is necessary to rule out a direct causal relationship between P and B.[7] We are therefore pushed to adopt the "common cause" model, shown in figure 5.8.

Let us now examine how the same conclusion may be reached on the basis of the Mentaculus. Once again, consider a particular, token episode of the problem. Suppose that in the actual world, we have a macro-history where:

- At t_0, no prediction has yet been made and the boxes haven't yet been filled.
- Shortly after t_1, it is predicted that the subject will one-box.
- At t_2, \$1M is deposited into the opaque box.
- At t_3, the subject decides to one-box.
- At t_4, the subject takes just the opaque box.
- At t_5, the subject goes home with her \$1M.

[7]So, for example, the only clues the subject is supposed to have about the prediction that has been made for her comes from the action that she eventually takes, and not from any external process connecting her to the prediction. Conversely, the stipulation that the subject makes her choice only *after* the boxes have already been filled is intended to rule out the B to P arrow, which would then have to be retrocausal.

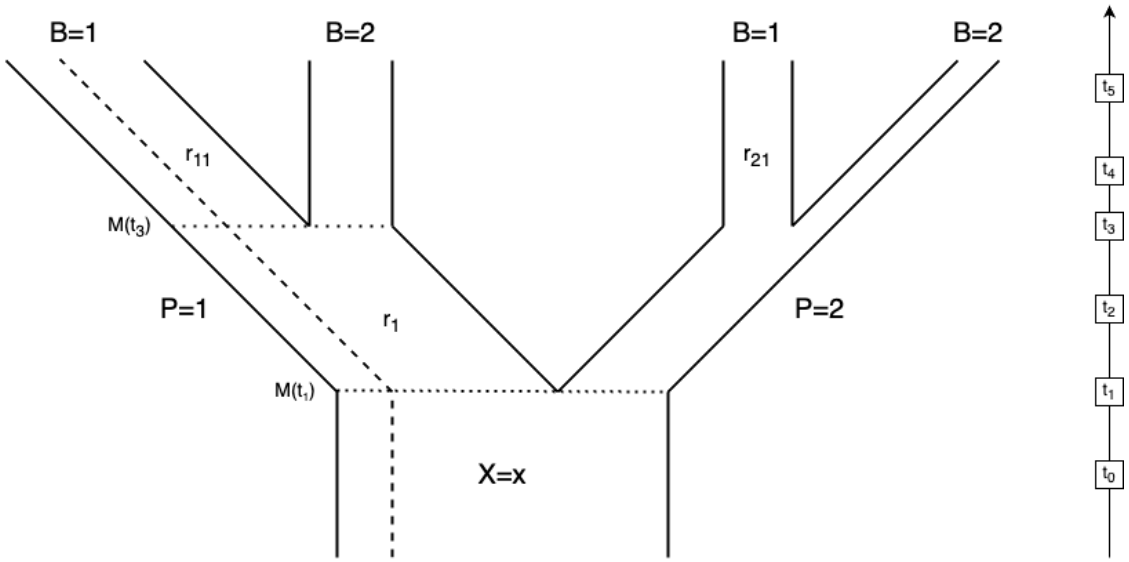

Figure 5.9: An instance of Newcomb's problem according to the Mentaculus.

A schematic diagram of the relevant part of the Mentaculus tree is shown in figure 5.9.

The main branch first splits into a $P = 1$ and a $P = 2$ sub-branch, with relative Mentaculus probability r_1 and $r_2 = 1 - r_1$ respectively. Each P-branch then slits into two further sub-branches, corresponding to $B = 1$ and $B = 2$. The Mentaculus probability of $B = 1$ conditional on the macrostate in which $P = 1$ is denoted r_{11}, and its complement $(P = 1, B = 2)$ is $r_{12} = 1 - r_{11}$. Similarly, the Mentaculus probability of $B = 1$ conditional on the different macrostate that would have obtained if it had been the case that $P = 2$ is denoted r_{21}, and its complement $(P = 2, B = 2)$ is denoted r_{22}. The dashed line represents the world's actual (micro-)trajectory, and the dotted lines represent the the (actual) macrostates at relevant branch points: $M(t_1)$ is the macrostate at the branch point between the $P = 1$ branch and the $P = 2$ branch, and $M(t_3)$ is the macrostate at the branch point between the $B = 1$ and $B = 2$ branches.[8]

We now impose the requirement that neither P causes B nor B causes P. If B doesn't cause P, then by the probabilistic intervention criterion (IC 3) in conjunction with the Mentaculus's semantics for counterfactuals (Counterfactuals 2), we must have

$$\Pr_{\mathcal{M}}(P = 1|B = 1, M(t^*)) = \Pr_{\mathcal{M}}(P = 1|B = 2, M(t^*)), \tag{5.16}$$

...where t^* is the branch-point time for the branching-off of $B = 2$ from $B = 1$. Provided that neither r_{11} nor r_{12} are negligibly small, that branch point time is t_3.

[8]Note that the branching of $B = 1$ and $B = 2$ occurs at t_3, the time of the subject's *decision*, rather than t_4, the time of the subject's *action*. Similarly, t_1 represents the branch-point time for the P-branches, which is expected to occur shortly *before* the time of the prediction.

We therefore find that $\Pr_{\mathcal{M}}(P = 1|B = 1, M(t_3)) = \Pr_{\mathcal{M}}(P = 1|B = 2, M(t_3)) = 1$, in satisfaction of this condition. So provided that neither r_{11} nor r_{12} are negligibly small, the branching structure of the Mentaculus ensures that B does not cause P.

Following the same line of reasoning, if P does not cause B then we must have:

$$\Pr_{\mathcal{M}}(B = 1|P = 1, M(t_1)) = \Pr_{\mathcal{M}}(B = 1|P = 2, M(t_1)) \tag{5.17}$$

$$\iff \Pr_{\mathcal{M}}(B = 2|P = 1, M(t_1)) = \Pr_{\mathcal{M}}(B = 2|P = 2, M(t_1)). \tag{5.18}$$

Here we have assumed that neither r_1 nor r_2 are negligibly small, and so the branch-point time for the branching-off of $P = 2$ from $P = 1$ is $t^* = t_1$. Thus we deduce that $r_{11} = r_{21} \Leftrightarrow r_{12} = r_{22}$; the ratio of the probability of $B = 1$ to $B = 2$ is the same conditional on both the $P = 1$ sub-branch and the $P = 2$ sub-branch. The intervention criterion tells us that the transition chances in this part of the Mentaculus tree must display precisely this probabilistic independence if we are to avoid the conclusion that P causes B.

I will now argue that the *forwards predictability* of the Mentaculus chances entails the existence of a variable X that acts as a common cause of P and B, and such that conditioning on this variable screens off the correlation between B and P (henceforth: a "screening-off common cause").

To begin with, let us re-write equations 5.17, 5.18, equivalently, in the form of a factorizability condition,

$$\Pr_{\mathcal{M}}(B = b, P = p|M(t_1)) = \Pr_{\mathcal{M}}(B = b|M(t_1)) \Pr_{\mathcal{M}}(P = p|M(t_1)). \tag{5.19}$$

Liouville's theorem (Goldstein et al. 2002, pp. 419–421) tells us that the phase-space measure of the sets of microstates satisfying $(B = b, P = p)$ (for any choice of $b, p \in \{0, 1\}$) remains constant over time. Thus, we can deduce from equation 5.19 that,

$$\Pr_{\mathcal{M}}(B = b, P = p|M(t_0)) = \Pr_{\mathcal{M}}(B = b|M(t_0)) \Pr_{\mathcal{M}}(P = p|M(t_0)). \tag{5.20}$$

This equation states that the "true" joint probability distribution (i.e., the Mentaculus chances) over the variables B and P, conditional on $M(t_0)$, factorizes into the product of its marginals. Now, the forwards predictability of macrodynamics (i.e., the fact that the exact distributional *forwards* time-evolution commutes with a Gibbs coarse-graining operation) tells us that these chances are predictable *merely from* $M(t_0)$ *itself*, even absent information about the precise form of the Mentaculus probability distribution over the $M(t_0)$ macrostate. Though the entire macrostate of the whole universe at t_0 must be sufficient, it is plausible that considerably less than that will do the trick; let us summarize the parts of the t_0 macrostate that are

relevant to the determination of the chances of B and P in the macro-variable X. Then, forwards predictability entails that we can make the identification,

$$\Pr_{\mathcal{M}}(B = b, P = p|M(t_0)) = P(B = b, P = p|X = x), \tag{5.21}$$

...where x is the value taken on by the variable X in $M(t_0)$. For the same reasons, we can make the analogous identifications for the marginal probabilities,

$$\Pr_{\mathcal{M}}(B = b|M(t_0)) = P(B = b|X = x) \tag{5.22}$$

$$\Pr_{\mathcal{M}}(P = p|M(t_0)) = P(P = p|X = x). \tag{5.23}$$

Let me linger briefly on the cognitive significance of these identifications. They assert that whenever certain macroscopic properties of the world's momentary physical state at t_0 are instantiated (bundled into the value of the macrovariable X), the underlying Mentaculus distribution over the macrostate instantiating those properties determines unique values for the chances of all future macrovariables, which are *independent of the detailed form of the Mentaculus distribution over the macrostate*. This is no kind of necessary truth; rather, it follows specifically from the Simple Dynamical Conjecture, taken in conjunction the postulated uniformity of the Mentaculus distribution over the Past Hypothesis at the initial time and (see section 3.2.2). And note that if it *weren't* true, then there would be no possibility of finding a macrovariable X that could achieve the identification of probabilities in equations 5.21–5.23. So while equations 5.21–5.23 in one way act as a definition of the macrovariable X, they also make the highly non-trivial assertion that such a macrovariable even *exists*.

The rest of the argument follows easily. The factorizability of the underlying Mentaculus chances (equation 5.20) together with the identifications of equations 5.21–5.23, entail the factorizability of the causal-models joint probability conditional on X, i.e.,

$$P(B = b, P = p|X = x) = P(B = b|X = x)P(P = p|X = x). \tag{5.24}$$

Thus, we must have,

$$P(B = b, P = p) = \sum_x P(X = x)P(B = b, P = p|X = x) \tag{5.25}$$

$$= \sum_x P(X = x)P(B = b|X = x)P(P = p|X = x). \tag{5.26}$$

The first equality (5.25) represents a decomposition of the overall joint probability $P(B = b, P = p)$ into components characterized by all the chance-determining

features of the $M(t_0)$ macrostate (summarized in the variable X), which may vary from case to case. Again, the forwards predictability of macrodynamics ensures us that for some such variable X, we can make the identification of each of these components $P(B = b, P = p|X = x)$ with a corresponding Mentaculus conditional probability. The second equality then follows from equation 5.20, which must hold true whenever P doesn't cause B.

The joint probability $P(B = b, P = p)$ therefore represents a kind of weighted average over many Mentaculus branches with the structure shown in figure 5.9, ranging over the different possible values for the variable X. Under the assumption that P doesn't cause B, each such branch must individually display a probabilistic *independence* of B conditional on P in order to satisfy the intervention criterion. However, an overall correlation between the two variables can emerge—and indeed, given the way the example is stipulated, must emerge—through the process of aggregating them together in the above manner.

In sum, then, the Mentaculus indeed supports the inference that *if* there is no direct causal relationship between P and B, and if nevertheless P and B are correlated variables (as we've stipulated), then there must exist a causal model of the situation involving a common-cause macrovariable X, conditioning on which screens off the correlation between P and B. This example can be taken to illustrate the general principle that the forwards predictability of Mentaculus-induced macrodynamics gives rise, in any particular situation, to the existence of a causal model satisfying the statistical constraint CM.

6
Conclusion

In the final remarks of his (2006) discussion of the Mentaculus theory of counterfactuals (referred to, in that paper, as the "SM-conditional"), Barry Loewer writes,

> [T]he account will prove its mettle if it can be connected to an account of causation. There is some hope of characterizing causation (or one concept of causation) in terms of counterfactual dependence and there is reason to think that the SM-conditional might play the role in a counterfactual analysis along along Lewisian lines.

This investigation follows up on Loewer's suggestion. Broadly speaking, my approach has indeed been to characterize causation in terms of counterfactual dependence, and then draw upon statistical mechanics (in the guise of the Mentaculus) to provide a naturalistic grounding for the relevant counterfactuals. However, while Loewer suggests exploring a counterfactual analysis of causation "along Lewisian lines", I have instead taken my cue from the formal methods of causal-statistical inference that are popular throughout the social, biological and information sciences, and in particular the framework of structural causal models. This framework already has its own implicit story to tell about the precise relationship between causal influence and counterfactual dependence, and it is this story that I tried to untangle in formulating the various versions of the intervention criterion.

In concluding, I will take the opportunity to revisit the three goals and motivations from Chapter 1 with which this project was conceived.

First, the basic motivating question—why do causes precede their effects in time?—has now been answered. The causal time-orientation is ultimately traceable to the *forwards-branching structure* of the Mentaculus. The time-asymmetry of this

branching structure results in a corresponding time-asymmetry of counterfactual dependence via the Mentaculus theory of counterfactuals, which explicitly exploits the branching structure. This, in turn, results in a time-orientation for the causal graph via the intervention criterion, which relates the edges in the causal graph—and, in particular, their *directions*—to these time-asymmetric counterfactuals. The details of this story are explained in section 5.2.

It is important to be clear about nature and scope of this result. As regards the *nature* of the causal time-orientation, we see that it is a *contingent fact* rather than a necessary truth. Neither the intervention criterion nor the Mentaculus theory of counterfactuals is necessarily bound to a time direction, as can be seen from their definitions. Together, they induce a causal time-orientation only in the context of the forwards-branching structure of the Mentaculus. This structure is a consequence of the low-entropy Past Hypothesis (in conjunction with the uniform distribution over the PH macrostate at the initial time); insofar as the low-entropy boundary condition on our universe is a contingent fact, so too is the forwards-branching structure, and so too, therefore, is the causal time-orientation.

As regards the *scope* of the causal time-orientation, it should not be forgotten that, according to the Mentaculus theory, the evaluation of most counterfactuals requires a limited amount of "backtracking" to the branch-point time, in order to lawfully realise the counterfactual antecedent $A(t)$. By choosing macro-variables located sufficiently closely together in time, or by enacting sufficiently drastic interventions on the world's macrostate at time t, this limited backtracking might suffice to "cover" the time t' of the consequent (in the case where $t' < t$), resulting in possible retrocausal arrows in the associated causal graph. The practical certainty that retrocausal arrows do not occur in realistic causal models such as those of figures 4.1 and 4.5 tells us something about the types of causal relationships that are likely to be of relevance to us as limited, spatially located human beings. Still, it is important to be clear that we cannot strictly say "causes *always* precede their effects in time", where the quantifier ranges over all imaginable interventions and possible consequences of those interventions; the truth about the causal time-orientation is a little more parochial than that.

Second, another goal of this investigation has been to offer a naturalistic justification for the various assumptions implicit in the use of formal causal models as a tool for causal reasoning. In particular, causal models generally (i) presuppose a directed, acyclic graph (DAG) structure for systems of causal relationships, and (ii) bake in a controversial commitment concerning the link between causation and statistics. The attempt to build a bridge between the causal models formalism

and the more foundational system of the Mentaculus is very much in the spirit of any other project of reduction between the models of a high-level theory and the models of a lower-level theory. Insofar as the models of the high-level theory work well to describe the world in a certain regime, there should in principle be an explanation for this success in terms of a reduction to the lower-level theory. Thus, insofar as certain aspects of the phenomena studied in the social, biological and information sciences conform to the structure of a formal causal model, there ought to be an explanation of this fact in terms of how those phenomena arise out of more fundamental phenomena in (for example) statistical physics.

Our investigation has reached certain conclusions on these points, too. The DAG structure was seen to be a simple logical consequence of time-orientation (see section 5.3). Though the DAG structure indeed *follows* from assuming a time orientation for the causal graph, I did not argue that it rests *solely* on an assumed time orientation. This would be an interesting area for future work: if it could be argued that the DAG structure does indeed rest solely on an assumed time orientation, then, insofar as the time-order criterion for the direction of causation admits of exceptions, so too might the DAG structure for the causal graph. This possibility raises a host of thorny issues that I did not have the space to tackle in this work.

As regards the more controversial link between causation and statistics, our investigation has ultimately ratified the presuppositions of the SCM formalism (which were carried over by fiat into the generalized ICM formalism), though the argument for this proved, perhaps appropriately, somewhat more elaborate. The key principle, which is itself a highly non-trivial idea in the foundations of statistical mechanics, is that the time evolution of a "simple" probability-density distribution on phase space may be computed (for the purposes of making macroscopic predictions) by regularly interspersing the dynamical time-evolution with a (Gibbs) coarse-graining operation. This is what Wallace (2023) calls the Simple Dynamical Conjecture. Combining this principle with the Mentaculus's simple cosmological distribution at the initial time (i.e., the uniform distribution over the PH macrostate), we deduce that the evolution of the world *forwards* in time is Markovian *even at the macroscopic level*. This is what always guarantees the existence of a causal model satisfying the Causal Markov Condition, and concomitantly underwrites the rule of product decomposition in all acyclic causal graphs.

Third, and finally, we can now look back at what we were able to achieve in this investigation and reflect on the effectiveness of the Mentaculus itself as a comprehensive system for naturalized metaphysics. I should begin by underscoring once again that the Mentaculus system, together with its theory of counterfactuals, appears

to be the convergent terminus of at least two largely independent philosophical traditions: on the one hand, we have the tradition, going back to Boltzmann, of thinking about time direction in light of the insights bequeathed to us by the physics of statistical mechanics; on the other hand, we have David Lewis's metaphysical edifice of laws, counterfactuals, and causation, which was self-professedly constructed in ignorance of the connection to statistical mechanics. Consequently, there is (to put it lightly) strong theoretical incentive to explore the Mentaculus, or something in its vicinity, as a promising candidate for a comprehensive metaphysical system.

But let us now judge it by its fruits. Despite not being explicitly designed for the job, the Mentaculus has by and large proved itself serviceable for grounding an account of causation. The clearest demonstration of this was given in section 5.1, where a full causal model for the Simpson's paradox case was reconstructed on the basis of the Mentaculus alone. In this sense, we can indeed say that the Mentaculus has, in Loewer's phrase, "proved its mettle".

One of the major *difficulties* that we encountered in connecting up the Mentaculus to an account of causation concerns the question of determinism. The determinism here at issue is not that of the underlying dynamical laws (which were taken as deterministic throughout), but rather that of the emergent causal relationships themselves. Do we say that causes operate to determine the *values* of their effects, or do we say instead that they operate to determine the *chances*? As we have seen, the framework of structural causal models presupposes the former, whereas the Mentaculus is seemingly constrained to work only in terms of the latter. We managed to circumvent this difficulty by constructing a generalized causal-models formalism—the formalism of *indeterministic* causal models (ICMs). Nevertheless, it is reasonable to wonder whether the apparent incompatibility of the Mentaculus with the usual formalism of *structural* causal models should already be a cause for concern.

A variety of different responses to this concern are possible, depending on one's philosophical inclinations. One sort of response places significant weight on the fact that many (most?) natural language statements of causal and counterfactual relationships take a definite form. (E.g., "My knocking the glass over caused it to fall to the floor" and "if I hadn't cut him off in the car, he would never have met his future wife".) For a philosopher moved by this sort of consideration, the fact that the Mentaculus is forced to press such definite claims into an awkward probabilistic form counts as evidence against it. Rather than adjusting the causal-models formalism to handle the probabilistic form of counterfactual conditional, such a philosopher might be more inclined to modify instead the Mentaculus theory of counterfactuals to handle the definite form.

Another sort of response downplays the evidence of natural language in favour of more theoretical considerations. In particular, the considerations presented in section 2.2.1 suggest that, if indeed the familiar *definite* form of causal/counterfactual statement is to be taken at face value, then almost all such counterfactuals come out false. If instead we interpret such natural language statements as gesturing towards a different, probabilistic sort of conditional of the form treated by the Mentaculus (e.g., "my knocking the glass over raised the probability of its falling to the floor close to one"), then we can avoid the threatened slide into an error theory about vast swathes of our causal/counterfactual discourse.

I am more sympathetic to this second sort of response, and have proceeded in this project on that basis. However, it would be disingenuous to pass over this point too quickly, so let me reiterate: if the Mentaculus is to provide the foundation for an account of causation, then causation must generically be conceived as probabilistic in nature, and the causal-models formalism must be adapted accordingly. I do not rule out the (epistemic) possibility that arguments from other quarters—whether they be ones I have as of yet overlooked or ones that are still waiting to be formulated—could still make the case for deterministic causation overwhelming. In this eventuality, the Mentaculus might need to face a reckoning.

Bibliography

Albert, David Z. (2000). *Time and Chance*. Harvard University Press.

Arntzenius, Frank (1992). "The Common Cause Principle". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2, pp. 227–237.

Boltzmann, Ludwig (1995). *Lectures on Gas Theory*. Dover Books on Physics. Dover Publications.

Brown, Harvey R., Wayne Myrvold, and Jos Uffink (2009). "Boltzmann's H-theorem, its discontents, and the birth of statistical mechanics". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 40.2, pp. 174–191.

Cartwright, Nancy (2002). "Against Modularity, the Causal Markov Condition, and Any Link between the Two: Comments on Hausman and Woodward". In: *The British Journal for the Philosophy of Science* 53.3, pp. 411–453.

Earman, John (1992). *World Enough and Space-Time: Absolute Vs. Relational Theories of Space and Time*. Bradford Books. MIT Press.

— (2006). "The "Past Hypothesis": Not Even False". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37.3, pp. 399–430.

Elga, Adam (2001). "Statistical Mechanics and the Asymmetry of Counterfactual Dependence". In: *Philosophy of Science* 68.S3, S313–S324.

Farr, Matt (2020). "C-theories of Time: On the Adirectionality of Time". In: *Philosophy Compass* 15.12, e12714.

Fine, Kit (1975). "Critical Notice". In: *Mind* 84.335, pp. 451–458.

Forster, Malcolm R. (1988). "Sober's Principle of Common Cause and the Problem of Comparing Incomplete Hypotheses". In: *Philosophy of Science* 55.4, pp. 538–559.

Frigg, Roman (2009). "Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics". In: *Philosophy of Science* 76.5, pp. 997–1008.

Glymour, Clark and Frank Wimberly (2007). "Actual Causes and Thought Experiments". In: *Causation and Explanation*. Ed. by J. K. Campbell, M. O'Rourke, and H. S. Silverstein. MIT Press, pp. 4–43.

Goldstein, H., C.P. Poole, and J.L. Safko (2002). *Classical Mechanics*. Addison Wesley.

Hájek, Alan (2020). "Counterfactual Scepticism and Antecedent-Contextualism". In: *Synthese* 199.1-2, pp. 637–659.

Halpern, Joseph Y. (2016). *Actual Causality*. The MIT Press.

Hausman, Daniel M. (1998). *Causal Asymmetries*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press.

Hausman, Daniel M. and James Woodward (1999). "Independence, Invariance and the Causal Markov Condition". In: *The British Journal for the Philosophy of Science* 50.4, pp. 521–583.

Horwich, Paul (1987). *Asymmetries In Time: Problems in the Philosophy of Science*. MIT Press.

Humphreys, Paul (1989). *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*. Princeton University Press.

Lewis, David (1973). *Counterfactuals*. Malden, Mass.: Blackwell.

— (1979). "Counterfactual Dependence and Time's Arrow". In: *Noûs* 13.4, pp. 455–476.

— (1981). "'Why Ain'cha Rich?'" In: *Noûs* 15.3, pp. 377–380.

Loewer, Barry (2006). "Counterfactuals and the Second Law". In: *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Ed. by Huw Price and Richard Corry. Oxford University Press.

— (2020). "The Mentaculus Vision". In: *Statistical Mechanics and Scientific Explanation*. Chap. 1, pp. 3–29.

Loewer, Barry (2023). *New Probabilistic Account of Counterfactuals*. URL: https://philsci-archive.pitt.edu/22971/.

Maudlin, Tim (2002). "Remarks on the Passing of Time". In: *Proceedings of the Aristotelian Society* 102, pp. 259–274.

McTaggart, J. Ellis (1908). "The Unreality of Time". In: *Mind* 17.68, pp. 457–474.

Mitchell, Melanie (2009). *Complexity : A Guided Tour*. Oxford University Press.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher. Reidel, pp. 114–146.

Papineau, David (1989). "Pure, Mixed, and Spurious Probabilities and Their Significance for a Reductionist Theory of Causation". In: *Minnesota Studies in the Philosophy of Science* 13, pp. 307–348.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Price, Huw (1997). *Time's Arrow & Archimedes' Point: New Directions for the Physics of Time*. Oxford University Press.

Reichenbach, Hans and Maria Reichenbach (1999). *The Direction of Time*. Dover Books on Physics. Dover Publications.

Salmon, Wesley C. (1985). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993). *Causation, Prediction, and Search*. The MIT Press.

Strevens, Michael (2011). "Probability Out of Determinism". In: *Probabilities in Physics*. Ed. by Claus Beisbart and Stephan Hartmann. Oxford University Press, pp. 339–364.

Wallace, David (2010). "Gravity, Entropy, and Cosmology: in Search of Clarity". In: *The British Journal for the Philosophy of Science* 61.3, pp. 513–540.

— (2015). "The quantitative content of statistical mechanics". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 52, pp. 285–293.

— (2018). *The Necessity of Gibbsian Statistical Mechanics*. URL: https://philsci-archive.pitt.edu/15290/.

— (2023). "The Logic of the Past Hypothesis". In: *The Probability Map of the Universe: Essays on David Albert's* Time and Chance. Ed. by Barry Loewer, Brad Weslake, and Eric B. Winsberg. Harvard University Press, pp. 76–109.

Winsberg, Eric (2023). "The Metaphysical Foundations of Statistical Mechanics: On the Status of PROB and PH". In: *The Probability Map of the Universe: Essays on David Albert's* Time and Chance. Ed. by Barry Loewer, Brad Weslake, and Eric B. Winsberg. Harvard University Press, pp. 57–76.

Woodward, James (2004). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in the Philosophy of Science. Oxford University Press.