Published in *Noûs* 44, 2010, pp. 80 -111.

CAUSATION: DETERMINATION AND DIFFERENCE-MAKING

Boris Kment

In his classic discussion of causation, David Hume introduced two ideas that have shaped much of the philosophical debate on the topic since then. The first may be called

> *The determination idea*. Causes determine that their effect obtains. If the causes of *E* obtain, then *E* inevitably obtains as well.

Hume seems to have taken it for granted that we ordinarily associate the idea of causation with that of a special tie or link between the causes and the effect in virtue of which the causes determine that the effect obtains, a 'necessary connexion,' as Hume calls it. He famously maintained that this idea of a tie is not based on any impression of such a connection, and that the only thing in the objects that could have given rise to it (by way of generating an association in the mind) is the constant conjunction of certain types of matters of particular fact. That train of thought yielded his definition of causation as constant conjunction. Descendants of this account were popular for a long time. Recent versions typically add the thought that the regularity in question must obtain as a matter of law. That is to say, the obtaining of matters of particular fact that are relevantly similar to the causes *nomically determines* (*is nomically sufficient for*) the obtaining of some matter of particular fact that is relevantly similar to the effect. (This view can be combined with a non-Humean theory of lawhood, and thus be divorced from its Humean origins.)

The *other* important idea about causation is introduced by Hume without stage-setting or obvious connection to the rest of the text. In the *Enquiry*, at the end of the section that deals with causation, Hume states his regularity account thus:

> "… we may define 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*."

To the surprise of the reader, the passage continues:

> “Or, in other words, where, if the first object had not been, the second never had existed.”[1]

The second formulation introduces a new idea, by no means identical with the one expressed by the first definition. We may call it

> *The difference-making idea*. A cause makes a difference to whether its effect obtains: without it, the effect would not have obtained.

This idea, too, is undeniably central to our thinking about causation. That is most clearly manifested in the methods we use to evaluate causal claims. The difference-making idea underlies John Stuart Mill’s method of difference (Mill 1956, Bk. III, ch. VIII, sct. 2), as the name of the method already suggests. Consider how Mill’s procedure is applied in the controlled experiments of science. As you manipulate the independent variable while controlling for other relevant factors, the value of the dependent variable changes. So, the value of the first variable makes a difference to the value of the second. There must therefore be a causal connection. Controlled experiments are frequently hailed as the foremost procedure for acquiring causal knowledge in the sciences. If that is right, then the difference-making idea forms the backbone of the scientific study of causal relationships. The central role of the difference-making idea in causal investigation is also attested by the fact that our causal judgments are so often guided by counterfactuals. If you want to know whether Fred’s tactless remark on Friday caused his fight with Susie on Sunday, what could be more natural than to ask whether the fight would have taken place without the remark? And if you decide that there would have been no fight without the remark—i.e. that the remark made a difference to whether the fight occurred—then it is very compelling to conclude that the remark was a cause of the fight.

The intuitions that support the determination and difference-making ideas are not about the way causation is to be *analyzed*. All intuition tells us is that there is a close connection of some kind between nomic determination and causation, and between causation and difference-making. Nonetheless, philosophers have long been trying to exploit these connections to give reductive accounts of causation. Initially, the determination idea was in the limelight, often in the form of minimal-sufficiency

[1] Hume 1995, p. 87.

accounts like the one propounded by Mackie.[2] This approach faced a number of formidable obstacles, which could never be resolved satisfactorily. By the early seventies it seemed time to try out something else, and after Lewis had published his influential paper on causation in 1973,[3] the attention of many philosophers shifted to the attempt to build an analysis of causation on the difference-making idea, which was most often articulated in counterfactual terms. This project confronts considerable problems of its own. Simple counterfactual dependence between distinct matters of particular fact is not necessary for causation, and arguably not sufficient either (see sections 1.1 and 1.2), so that a counterfactual analysis cannot simply equate causation with such counterfactual dependence, but must define causation as some *complex pattern* of counterfactual dependencies. There have been numerous attempts to do that in a way that gets the extension of causation right, but, in my opinion, they met with only limited success. Moreover, there are some reasons for thinking that a correct account of counterfactuals requires causal notions, so that causation cannot in turn be analyzed in counterfactual terms without circularity (see section 1.3).

Even if both attempts to analyze causation fail, it seems plausible that the concepts of nomic determination and difference-making figure prominently in our causal thinking, and any good philosophical theory of causation ought to explain that fact. But there is also another, closely related task that a theory of causation faces in this area. It should seem very puzzling that the determination and difference-making ideas are *both* so compelling. For, as many readers of Hume have remarked, the two ideas are quite different (contrary to what Hume suggests). To say that the causes together nomically determine their effects is to say that, given the laws, the causes are *jointly sufficient* for the effect. By contrast, to say that the effect would not have obtained if any of causes had not obtained is to say that causes are *individually necessary in the circumstances* for the effect. What could have possessed Hume to define one and the same notion in *both* of these ways? And how can our thinking about causation be governed by two ideas that are so different? A good theory of causation should solve this riddle by telling us if and how the two ideas are connected.

[2] See, e.g., Mackie 1974, Hall 2004a, sct. 7, Strevens 2007.

[3] Lewis 1986b.

My focus will be on explaining the role of the notion of difference-making in causal thinking, and on its connection to the determination idea. The counterfactual analysis of causation gives the most straightforward explanation of the importance of the concept of difference-making in causal inquiry: causation *consists in* a certain pattern of counterfactual dependencies between distinct matters of particular fact. Hence, to ask whether *C* caused *E* is simply to ask whether certain counterfactuals hold. I will offer an altogether different explanation. Patterns of difference-making, like those we study in scientific experiments and counterfactual reasoning, are not what *makes* causal claims true. They merely provide a useful *test* for causal claims. Moreover, I will argue that what justifies us in using them to test causal claims is the determination idea. That is how the determination and difference-making ideas are connected. This account, as we will see, predicts and explains all the phenomena that present difficulties for counterfactual analyses. The very findings that threaten to refute the counterfactual analysis therefore confirm my account.

In section 1, I will describe some background facts about the difference-making idea, and in section 2 I will do the same for the determination idea. That will set the stage for the exposition of my account in the remainder of the paper.

1. The difference-making idea

I will begin by considering the limitations of the difference-making idea (understood in counterfactual terms) and the phenomena that stand in the way of turning it into an analysis of causation: the extensional difference between causation and counterfactual dependence, and the role of causal notions in the truth-conditions of counterfactuals. These are among the data that my account is intended to explain.

1.1 *Counterfactual dependence and deterministic causation*

There are two types of example that show that counterfactual dependence between distinct matters of particular fact is not necessary for causation under determinism.[4] *Firstly*, cases of preemption: *E* is caused by *C*, but there is a second potential cause of *E*, which is prevented by *C* from causing *E*. If *C* had not obtained, then the backup cause

[4] For more detailed expositions of the over-determination and preemption problems, see Lewis 1986b (including the postscripts), Menzies 1989a, and Schaffer 2004a.

would have taken over and caused *E*. To illustrate: Susie and Billy plan to smash a certain bottle with bricks. Susie is there first. When Billy arrives, he sees Susie throw her brick and decides not to throw his. Susie's throw causes the bottle to shatter. But the shattering does not counterfactually depend on her throw. If she had not thrown her brick, Billy would have thrown his, and the bottle would have shattered anyway. *Secondly*, there are cases of 'symmetrical over-determination,' in which two independent causal chains lead up to the same effect and both run to completion. Suppose that Fred's and Susie's bricks arrive at the bottle at the same time, each of them causing sufficient damage to shatter the bottle. It seems natural to regard the collision of each brick with the bottle as a cause of the shattering. But the shattering does not counterfactually depend on either collision. If one of them had not occurred, the other would have done the work.[5]

Preemption and over-determination cases are well-known, and it is consequently widely accepted that counterfactual dependence is not necessary for causation.[6] There are also good reasons for doubting that counterfactual dependence between distinct matters of particular fact is *sufficient* for causation. It may be a *near*-sufficient, but there are some recherché examples of counterfactual dependence without causation.

Some stage setting is required before I can discuss these cases. On the standard view, a counterfactual is true just in case its consequent is true in the "closest possible antecedent-worlds," i.e. in those possible worlds where the antecedent is true and which otherwise resemble our world as closely as possible.[7] The crucial question is what rules

[5] Some philosophers (e.g., Lewis 1986b, postscripts) would deny that each of the bottle-brick collisions is a cause of the shattering. But to me (and to many others) it seems intuitively plausible that each collision is a cause, and for the sake of determinateness I will assume that that is so. Nothing of importance hinges on this assumption. If you do not agree, you can simply ignore my future uses of over-determination examples as cases of causation without counterfactual dependence. In that case, you should still take preemption cases to show that counterfactual dependence is not necessary for causation.

[6] Counterfactual theorists of causation have shown no lack of ingenuity in reacting to cases of preemption and over-determination. (See, e.g., Lewis 1986b (including postscripts), Menzies 1989a, McDermott 1995, Ramachandran 1997, Lewis 2004, Yablo 2004.) Some of them have tried to solve the problem by appealing to the idea that causation is transitive, others by appealing to its supposedly intrinsic nature, or to the thought that causes and effect need to be spatio-temporally connected by continuous causal chains (to mention just some of the strategies). More recently, philosophers using the framework of causal models have proposed a number of other ways of dealing with over-determination and preemption problems. (See, e.g., Pearl 2000, Hitchcock 2001, Woodward 2003, Halpern and Pearl 2005, and Hall 2007 for good discussions.) It is beyond the scope of this paper to review these strategies, not least because any attempt to do so would quickly get bogged down in trench warfare (to borrow Tim Maudlin's phrase). Suffice it to say that the considerable complications and difficulties that counterfactual analyses face provide plenty of justification for exploring new routes.

[7] This theoretical framework is due to Stalnaker (1968) and Lewis (1973). Other significant work done in this framework includes Jackson (1977), Bennett (1984), and Lewis (1986c), in addition to the writings mentioned later on in this paper. The standard Stalnaker-Lewis account has the problematic consequence

determine which antecedent-worlds count as the closest. Consider a counterfactual about matters of particular local fact, such as "If Nixon had pressed the button at time *t*, there would have been a nuclear catastrophe." Most philosophers would agree on two data points: if Nixon had pressed the button at *t*, then up to *t* things would have been pretty much the way they actually were; after *t*, events would have unfolded in accordance with the actual laws, generating the catastrophic consequence mentioned in the consequent. So, there are two desiderata for the closest antecedent-worlds: match in matters of particular fact up to the antecedent-time, and conformity to the laws of our world.

There is one complication, however. Under determinism, any initial segment of the actual world's history, together with the laws, determines that Nixon does *not* press the button. Hence, there is no antecedent-world that meets both desiderata perfectly. Antecedent-worlds that conform perfectly to our laws must be unlike our world throughout the pre-antecedent time. And antecedent-worlds that are exactly like our world right until the antecedent-time must feature a big and conspicuous violation of the actual laws, a big 'miracle,' as Lewis calls it. (Suppose that the button is on the second floor of the White House, and that in the actual world Nixon was on the first floor at *t*. In antecedent-worlds that are like our world right until *t*, Nixon suddenly disappears from the first floor and reappears on the second, with his finger pressing the button.) As Lewis has shown, however, there are also antecedent-worlds that are *almost* exactly like our world until the antecedent-time and which do not feature any big and conspicuous miracles. They diverge from our world shortly *before* the antecedent-time: a minute before *t*, Nixon decides to walk upstairs and press the button. Under determinism, this still requires a violation of the actual laws. But the violation can be small and inconspicuous. Maybe some extra neurons miraculously fire in Nixon's brain. Lewis maintains that these worlds provide the best trade-off between the two desiderata of match up to the antecedent-time and conformity to the actual laws. I agree (and present an argument for this view in my (2006a)).

that all counterfactuals with impossible antecedents are true. I think that the best way of resolving this problem is to let impossible worlds to figure in the theory of counterfactuals alongside possible worlds (see Nolan (1997) and my (2006b)). For the purposes of this paper, however, the more familiar account in terms of possible worlds will work just as well, and we can ignore impossible worlds. In what follows I will use 'world' as synonymous with 'possible world.'

The closest antecedent-worlds, then, diverge from our world before the antecedent-time (a phenomenon known as "backtracking").[8] But that presents a problem for counterfactual accounts of causation, as is shown by examples described by Jonathan Bennett.[9] Here is an example of the same kind, which is due to Peter Lipton.[10] A gigantic hydrogen bomb explodes in Detroit at noon. The pressure wave spreads outwards and destroys Ann Arbor at 12:03. What if the bomb had not destroyed Ann Arbor? The closest antecedent-worlds diverge from our world shortly before the antecedent-time, by a small and inconspicuous violation of the laws (under determinism) or without any violation (in certain indeterministic cases). Now, once the bomb has exploded, only a big and conspicuous miracle could prevent it from destroying Ann Arbor. The closest antecedent-worlds must therefore diverge *before* that and omit the explosion altogether. But in such worlds Detroit does not get destroyed either. Hence, if Ann Arbor had not been destroyed, then Detroit would not have been destroyed. Similarly, if Ann Arbor had not been destroyed, then Detroit would not have been in ruins the next day. These are cases of counterfactual dependence without causation.

1.2 *Counterfactual dependence and probabilistic causation*

In addition to the difficulties described in the previous section, there are some problems for counterfactual accounts of causation that are specific to the indeterministic case. If indeterminism is pervasive, so that it is almost always a matter of chance what happens, then it is rarely true that *X would not* have obtained if things had been different in a certain way. The most we can say is that, if things had been thus-and-so, then the chance of *X*'s obtaining would have been different in such-and-such ways. Effects therefore do not generally counterfactually depend on their probabilistic causes, and we cannot analyze indeterministic causation in terms of a pattern of counterfactual dependence between cause and effect. Instead, counterfactual theorists typically appeal to patterns of dependence that link the cause to the *chance* of the effect. The most common version of this account starts from the idea that causes raise the probabilities of their effects.

[8] Jackson (1977) disagrees. See Bennett (2003, sct. 79) for arguments against Jackson's position and in favor of Lewis'.

[9] Bennett 1984.

[10] Presented in a lecture at Cambridge University in 1997.

Without the cause, the effect would have been less likely. (This account has been developed further in several different ways by a number of philosophers.)[11]

Christopher Hitchcock characterized the idea underlying this view memorably as follows:

> "Various causes increase the probability of an effect by contributing to a 'probability pool.' Once the probability of the effect is determined, the dice are cast, and the event either occurs or it does not. The individual causes make no additional contribution to the outcome; they bring it about only via their contribution to the probability pool." (Hitchcock 2004, p. 407)

Unfortunately, it can be shown that indeterministic causation is *not* merely a matter of contributing to the probability pool. Consider a case due to Jonathan Schaffer (Schaffer 2000). Merlin casts a spell to turn the prince and the king into frogs at midnight, and Morgana casts a spell to turn the prince and the queen into frogs at midnight. Once one of these spells has been cast, its chance of success remains constant at 50% until midnight. Since the results of the two spells are stochastically independent, the two spells result in a chance of 75% that the prince will become a frog at midnight. At midnight the prince is transformed along with king, while the queen is not. The result proves that Merlin's spell worked, while Morgana's was ineffective. So, Merlin's spell is a cause of the prince's transmutation while Morgana's is not. Note, though, that their contributions to the probability pool of the prince's transformation are exactly the same. Each of them raised the probability from 50% to 75%. This shows that we cannot determine *c*'s role in bringing about *e* merely by looking at how *e*'s chance depends on *c*.

Schaffer's example illustrates the important difference between influencing *the chance* of some matter of particular fact and influencing *whether it obtains*. (Morgana's spell influenced the *chance* of the prince's transmutation—it raised it to 75%—but did not causally contribute to the *occurrence* of the transmutation.) Now, it seems plausible enough that counterfactuals about chances can be used to support claims about the causes *of these chances*. If we know that E would not have had chance p at time t if C had not obtained, then we have reasons for concluding that E had chance p at t (that $\mathrm{ch}_t(E) = p$, for short) at least in part because of C. Admittedly, the connection between counterfactuals about chances and causal claims about these chances is subject to the

[11] For a classic statement of this view, see Lewis (1986b), postscript B.

same limitations as the connection between counterfactuals and causal claims under determinism. *C* may be a preempting or over-determining cause of the fact that $\text{ch}_t(E) = p$, in which case this fact does not counterfactually depend on *C*. Or the fact that $\text{ch}_t(E) = p$ may counterfactually depend on *C* for backtracking reasons, in the absence of a causal connection. But it seems plausible that the connection between counterfactuals about chances and claims about the causes of these chances presents no problems that do not already arise for the connection between counterfactuals and causal claims under determinism. We should therefore expect that any counterfactual analysis of causation that can be made to work for the deterministic case can be extended to propositions about the causes of chances. But, given that indeterministic causation of *other* effects (i.e., effects that are not facts about chances) is not merely a matter of influencing chances, it is not obvious how to extend the account to such instances of causation. However, such an extension would be needed in order to obtain a unified counterfactual account that covers all cases of causation.

1.3 *Causal notions in the theory of counterfactuals*

Another word on counterfactuals. Under determinism, the picture described in section 1.1 is essentially adequate. The closest antecedent-worlds diverge from ours by a small miracle shortly before the antecedent-time, so that the antecedent comes out true. After that they evolve in accordance with the actual laws. The history up to the antecedent-time, together with the deterministic laws, determines what the rest of history looks like.

Under indeterminism the story is somewhat more complicated. The closest antecedent-worlds are, again, pretty much like our world until around the antecedent-time, and then diverge so as to make the antecedent true. Under indeterminism, that may not even require a violation of law. It may be enough that some random processes have different outcomes. After the antecedent-time the worlds evolve in accordance with the actual laws. But under indeterminism, the history up to the antecedent-time and the laws need not determine the rest of history. For there are different ways the post-antecedent chance processes can turn out. Some outcomes make the post-antecedent history more similar to the history of the actual world than others. That raises the question of whether, in addition to the two desiderata for closeness that we already considered (match until the antecedent-time and conformity to the actual laws), there is (under indeterminism) a third desideratum of post-antecedent similarity.

The answer is a qualified 'yes.' Some post-antecedent similarities matter, others do not. Consider a variant of an example due to Dorothy Edgington (2003).[12] You are about to watch an indeterministic lottery draw on television. Just before the draw, someone offers to sell you ticket number 17, but you decline. As it happens, ticket number 17 wins. It seems true to say 'If you had bought the ticket, you would have won,' but that presupposes that

> If you had bought ticket number 17, that ticket would still have won.

Contrast that with:

> If they had used a different machine in the draw, 17 would still have won.

That seems false. If they had used a different machine, then 17 might have won, or some other number might have won. It is not true that 17 *would* have won.

In the first case, we hold the outcome of the lottery draw fixed, in the second case we do not. It is in explaining this difference that causal notions are often brought into the theory of counterfactuals.[13] For the most plausible explanation is this. Your decision whether to buy the ticket is not causally connected to the outcome. That is why the outcome would have been just the same if you had made a different decision. The second example is different. The use of a particular lottery machine is part of the causal history of the outcome. You change that causal history when you replace the machine with another. That is why it contributes nothing to the closeness of an antecedent-world if the outcome is the same. The upshot: when we reason about how things would have been different if some matter of particular fact *C* had not obtained, match in post-antecedent matters between an antecedent-world and our world contributes to closeness if and only if these post-antecedent matters are causally independent of *C* in our world.[14] (For a more precise statement of this principle and further discussion of it, see my (2006a).)

That yields the following picture. Suppose that *A* is some actual matter of particular local fact that obtains at *t*. Under indeterminism, the closest ~*A*-worlds are just like our world until shortly before *t*. Then they diverge from our world (by a small miracle or

[12] Similar examples can be found, e.g., in Tichý 1976, and Bennett 2003, ch. 15.

[13] See Adams 1975, ch. IV, sct. 8, in particular pp. 132f., Edgington 1995, sct. 4.4, 2003, Mårtensson 1999, Bennett 2003, ch. 15, Schaffer 2004b, Hiddleston 2005, Kment 2006a, and Wasserman 2006.

[14] By '*E* causally depends on *C*,' I mean that *C* stands in the ancestral relation of causation to *E*. (Whether that entails that *C* is a cause of *E* depends on whether causation is transitive. That is a controversial issue. While some philosophers defend the assumption of transitivity (e.g. Lewis in his 2004), others have been convinced by apparent counterexamples to reject the assumption. (For useful discussions of this issue, see, e.g., McDermott 1995, Hall 2004b). I will remain neutral on that issue in this paper.)

without miracle[15]) so that ~*A* comes out true.[16] After that they unfold in accordance with the actual laws of nature. And they maximize match in those post-antecedent matters that are (in the actual world) causally independent of *A*. On this view, causal notions are needed in an account of counterfactuals, which precludes a reductive account of causation in counterfactual terms.

2. The determination idea

On the account I will advocate, our use of patterns of difference-making as a guide to the causal facts rests on our acceptance of the determination idea. That idea comes in different versions depending on whether we assume the truth or falsity of determinism (by which I will here understand the thesis that any possible world that matches ours at one time and which conforms to the actual laws is like our world at every time). Under determinism, the simplest version of the determination idea is the thesis that

(D/d) The set containing all and only *E*'s causes nomically determines (is nomically sufficient for) *E*,[17, 18, 19]

[15] Even under indeterminism a small miracle *may* be required for perfect match until shortly before the antecedent-time. If the world is indeterministic, then there can be forks, that is, cases where the outcome of a chance process contributes to determining which of several futures will be realized. But the thesis of indeterminism entails nothing about the frequency of forks, and it leaves open the possibility that they are extraordinarily rare. It may be that the latest relevant fork is located a long time before the antecedent-time, so that any antecedent-world that is perfectly like our world until shortly before the antecedent-time contains a violation of the actual laws of nature.

[16] In special cases, the closest antecedent-worlds may even diverge from the actual world *at* the antecedent-time, namely if no more than a small miracle at that time is required to make the antecedent come out true (that may be true in certain cases where the antecedent is about microphysical matters).

[17] '(D/d)' stands for 'determination idea/deterministic version.'

[18] (D/d) requires a qualification. There may be matters of particular fact that have no causes at all (e.g., those that obtain at the very beginning of the history of the world, if there are such). The set of causes of such an uncaused matter of particular fact *E* is the empty set, and the empty set does not nomically determine *E*. (D/d) must therefore be restricted to those matters of particular fact that have causes. I will leave this restriction implicit from now on.

[19] (D/d) could be read as a thesis about type causation or about token causation. In this paper, my focus is on token causation (though I suspect that much of my account also applies, *mutatis mutandis*, to uses of the counterfactual test to establish claims about type causation), and I understand (D/d) as a thesis about token causes. Different philosophers have different views about the kinds of entities that are token causes. Some think of them as events, others as facts, states of affairs, situations, event aspects or property instances (see, e.g., Davidson 1980, Kim 1973, Lewis 1986d, Bennett 1988, Menzies 1989b, Mellor 1995, 2004, Paul 2004). How we need to spell out the determination idea about token causation depends to some extent on the view we take on this issue. I think that causal relata can belong to different ontological categories, including the categories of events and facts. On my preferred understanding of (D/d) as formulated here, it is a thesis about token causation between *facts*: the fact that is the effect obtains in all possible worlds that

where for present purposes we can simply understand this as the thesis that *E* obtains in all possible worlds where all of *E*'s actual causes obtain and which conform to the actual laws of nature.[20] In contrast to the difference-making idea, (D/d) is a thesis, not about the individual causes of *E*, but about *the set of all causes of E*. (Individual causes of *E* may not nomically determine *E*; but all the causes of *E* taken together do.)

It seems plausible that principle (D/d) is true under determinism. Suppose that Susie throws a rock at a window and shatters it. And assume that determinism is true. Consider all the causes of the window shattering, *including omissions*. These causes include Susie's throw, the position and molecular structure of the window, etc. They also include the absence of any factors that could interfere with the shattering, such as obstacles in the path of the flying rock, strong winds that could blow the rock off its path, bystanders trying to catch the rock, and so on. Complete this list of causes, and you get a set of factors that nomically determine the breaking of the window. In every possible world that conforms to the actual laws and in which all of these causes obtain, the window shatters.

The determination idea as formulated in (D/d) does not tell us that we can give an analysis of the concept of causation, or a real analysis of the relation of causation, in

contains the facts that are the causes and which conform to the actual laws. In this paper I will try to show that this thesis underlies our use of counterfactuals to test claims about token causation between facts.

There is also, I think, a version of the determination idea that applies to type causes. In the case of causation between facts, type causes and type effects are *type facts*. A lengthy discussion would be necessary to give a satisfactory philosophical explanation of the concept of a type fact. But the intuitive idea underlying this notion is easy to grasp. As a very rough first shot (which requires a number of revisions and refinements that cannot be provided here), we can say that two token facts instantiate the same type fact just in case they have the same nomically relevant features, for some suitable notion of nomic relevance. (For example, if *f* is the fact that there is an explosion at time *t* in place *p*, and *g* is the fact that there is an explosion at time *t** in place *p**, then *f* and *g* may instantiate the same type fact.) I also think that it is a necessary feature of a token fact that it instantiates the specific type fact it does. (For example, it is a necessary feature of the fact *f* that it is the fact that there is an explosion at a certain time and place.) Now suppose that certain tokens *x*, *y*, *z* of the type facts *X*, *Y* and *Z* obtain on a certain occasion, and are followed by a token *e* of type fact *E*. And suppose that *x*, *y* and *z* include all the causes of *e*. Then according to (D/d), *x*, *y* and *z* together nomically determine *e*. Given that *e* is necessarily an *E* token, this entails that *x*, *y* and *z* together nomically determine the proposition that an *E* token obtains. Now, any *X*, *Y* and *Z* tokens *x**, *y** and *z** that obtain together on another occasion have the same nomically relevant features as *x*, *y* and *z*. Hence, given that *x*, *y* and *z* nomically determine that an *E* token obtains, *x**, *y** and *z** do so as well. So, we obtain a general principle that runs very roughly as follows:

(5) If the tokens $x_1, x_2, \ldots, x_n$ of type facts $X_1, X_2, \ldots, X_n$ include all the causes of a certain *E* token, and $y_1, y_2, \ldots, y_n$ are tokens of $X_1, X_2, \ldots, X_n$ that obtain together on another occasion, then $y_1, y_2, \ldots, y_n$ also nomically determine the obtaining of an *E* token.

[20] I think that modal concepts are not really needed to define nomic determination, but that the notion can instead be explained in terms of narrowly logical entailment. (I also think that that way of spelling out the determination idea is preferable for certain theoretical reasons.) But for the purposes of this paper, the simpler modal definition will do.

terms of nomic determination. In fact, (D/d) says nothing at all about whether causation, or the concept thereof, can be analyzed at all. It does not even offer necessary and sufficient conditions for causation. All it does is to state a *necessary* condition for a set to contain all the causes of *E* under determinism: the set must nomically determine *E*. There are many sets that meet this condition, and (D/d) does not tell us which of these sets is the set of *E*'s causes.

Determination analyses of (the concept of) causation under determinism typically start from the idea that, roughly speaking, a cause of *E* is a member of some set *S* of matters of particular fact that is minimally nomically sufficient for *E* (i.e., a set *S* that is nomically sufficient for *E* and none of whose proper subsets are nomically sufficient for *E*) and which does not entail *E*. The main problems for the view arise from the fact that that is not in fact a *sufficient* condition for causation. *C* can belong to a set of matters of particular fact that is minimally nomically sufficient for *E* and does not entail *E*, even if *C* is not a cause of *E*, but an effect of *E*, or a preempted potential cause of *E*, or if *C* and *E* have a common cause. But none of these counterexamples cast any doubt on the claim that (under determinism) it is a *necessary* condition for a set to contain all causes of *E* that the set nomically determines *E*. So, even if the determination *analysis* falters on the aforementioned problem cases, they provide no reason for doubting the truth of what I called the 'determination *idea*,' i.e. (D/d).

In fact, even if nomic-determination analyses fail, it may still be a *necessary* truth that (D/d) holds under determinism, e.g. because it is an essential property of the relation of causation that (D/d) holds under determinism. For the claim that the truth of (D/d) is essential to causation does not entail that causation can be analyzed in terms of nomic determination.[21] Compare: It may be essential to Fred to be human and to have originated from a certain sperm and egg, but that does not entail that we can give a real analysis (or real definition) of Fred—i.e., roughly speaking, an account of what it is to be Fred—in terms of some combination of his species, origin, and other conditions. (A real definition needs to state conditions that are necessary and sufficient for a thing in another possible

[21] I owe this point to Gideon Rosen.

world to be Fred.[22] Being human and originating from certain gametes is not sufficient, for Fred could have had an identical twin. And it is unclear what conditions to add to these two properties to obtain non-trivial necessary and sufficient conditions.) Similarly, it may be essential to Newtonian mass that its behavior is governed by certain laws, and it may yet be impossible to define mass by appeal to these laws. (Obeying these laws is not obviously sufficient for something in another possible world to be mass. And what conditions could we add to obtain necessary and sufficient conditions?)

Similarly, even if the determination analysis of the concept of causation fails, it may still be analytic on the concept of causation that (D/d) holds under determinism. Compare: It may be analytic on the concept of some normative property *F*-ness that the distribution of *F*-ness supervenes on the descriptive facts, even if this supervenience principle cannot be used to analyze the concept of *F*-ness. Similarly, it may be analytic on the concept of knowledge that X knows that *p* only if X believes truly that *p*, but it does not follow that the concept of knowledge is definable in terms of a combination of true belief and other conditions. (Attempts at such definition have not met with much success.)

I do not need to take a stand on whether (D/d) is necessary, or an essential or analytic truth, or on whether (the concept of) causation can be analyzed in terms of nomic determination, or in any other way. I do not even need to endorse the claim that (D/d) is true without exception. All I will assume is that our thinking about deterministic causation often relies on (D/d). That seems plausible. Suppose that I made a certain type of cake on two different occasions. One time it was delicious, the second time it was chalky and unappealing. Then it seems very tempting to say: I must have done something the second time that I didn't do the first time, and which made the second cake taste chalky. In other words, I conclude from the fact that the two cakes taste different that the factors that are causally responsible for the taste of the first cake must be somewhat different from those responsible for the taste of the second cake. That is the

[22] I am not assuming that it is *sufficient* for the correctness of a real definition that it states such necessary and sufficient conditions, merely that it is *necessary* for correctness.

contrapositive of the principle that, if all the causes are the same, then the effect must be the same. And that, in turn, looks like an application of the determination idea.[23]

Principle (D/d) can be strengthened. Suppose that *E* is some matter of particular fact that obtains at time t_E, and that *t* is earlier than t_E. Then it seems plausible that, under determinism,

(D/d*) Those causes of *E* that obtain no later than *t* jointly nomically determine *E*.

After all, under determinism any initial segment of our world's history contains sets of factors that nomically determine *E*. (D/d*) simply says that the set of causes of *E* that obtain in this initial segment includes a complete set of such nomic determiners of *E*. There is an even stronger version of the determination idea that also has a lot of intuitive force. Under determinism, if *E* obtains at t_E and *t* is earlier than t_E, then

(D/d**) Those causes of *E* that obtain at *t* jointly nomically determine *E*.

Under determinism the state of the world at any given time *t* before t_E contains sets of factors that nomically determine *E*. (D/d**) simply says that the set of all the causes of *E* that obtain at *t* includes a complete set of such nomic determiners of *E*.

(D/d*) and (D/d**) are not essential or analytic truths. The latter two principles can fail in cases of backwards causation or action at a temporal distance. But in a world like ours, these phenomena are at best extremely rare, and quite possibly non-existent (or so we typically think), and we can and do ignore these possibilities in all ordinary cases. So, our reasoning about deterministic causation can ordinarily proceed on the assumption that (D/d*) and (D/d**) are true.

Under indeterminism, it obviously need not be true that the causes of *E* nomically determine *E*. Hence, (D/d) does not hold. There is, however, an intuitively plausible indeterministic version of the determination idea that is restricted to the causes of one special kind of fact, namely facts about chances. Let *E* be some matter of particular fact obtaining at t_E, let *t* be some time before t_E, and let '$\mathrm{ch}_t(E)$' again stand for the chance of *E* at *t*. Suppose that $\mathrm{ch}_t(\mathrm{E}) = p$. Then,

(D/i) The causes of the fact that $\mathrm{ch}_t(E) = p$ jointly nomically determine that $\mathrm{ch}_t(E) = p$.[24]

[23] To be more precise, this reasoning process applies, not principle (D/d), but the closely related principle (5) of footnote 19. The taste of the first cake and that of the second cake instantiate different type facts. Hence, the factors that are responsible for the taste of the first cake cannot instantiate exactly the same types as the factors responsible for the taste of the second cake.

(D/i) seems plausible. If *E* has a certain chance at *t*, then there must be some matters of particular fact that causally determine that *E* has this chance at *t*.

(D/i) is not to be confused with

~~(D/i)~~ The causes of *E* jointly nomically determine that $\mathrm{ch}_t(E) = p$.

(Note the disanalogy between ~~(D/i)~~ on the one hand and (D/d) and (D/i) on the other. Both (D/d) and (D/i) say that certain effects are nomically determined by *their* causes. By contrast, ~~(D/i)~~ says that the fact that $\mathrm{ch}_t(E) = p$ is nomically determined by the causes of *something else*, namely by the causes of *E*.) ~~(D/i)~~ is false, as can be shown using Schaffer's example discussed in section 1.2. The causes of the prince's metamorphosis include Merlin's spell, but not Morgana's. Since they do not include Morgana's spell, they are not jointly nomically sufficient for the fact that the prince's transformation had a chance of 75%. Note that the same example does not refute principle (D/i). While Morgana's spell is not a cause of the prince's transformation, it *is* a cause of the fact that the transformation had a chance of 75% (and not 50%). Morgana's spell and Merlin's spell are both causes of the latter fact, and (together with certain background conditions) they nomically determine that fact.

(D/i) cannot be strengthened in the same way as (D/d). Let *E* be some matter of particular fact, let t_1 be some time before the time when *E* obtains, and assume that $\mathrm{ch}_{t1}(E) = p$. And let t_0 be some time before t_1. Then it need not be true that the causes of the fact that $\mathrm{ch}_{t1}(E) = p$ that obtain at t_0 nomically determine that $\mathrm{ch}_{t1}(E) = p$. It need not even be true that all the causes of the fact that $\mathrm{ch}_{t1}(E) = p$ that obtain *up to* t_0 nomically determine that $\mathrm{ch}_{t1}(E) = p$. For *E*'s chance at t_1 may be partly nomically determined by the outcomes of random processes occurring between t_0 and t_1.

3. The counterfactual test: the basic idea

3.1 *The method of elimination, the method of difference, and counterfactual reasoning*

Counterfactual dependence between distinct[25] matters of particular fact is not *quite* sufficient for causation—that much is shown by backtracking examples—but it seems to

[24] '(D/i)' stands for 'determination idea/indeterministic version.'

[25] It is not an easy task to say exactly which notion of distinctness is relevant here. Certainly, in order for the counterfactual dependence of *E* on *A* to show that there is a causal connection, *A* and *E* must be distinct in a sense that involves more than mere non-identity. I would guess that it also requires that the obtaining of *A* does not necessitate the obtaining of *E* or vice versa, and that it perhaps involves some form of mereological distinctness as well.

be *near-sufficient*. However, it is far from *necessary*, as preemption and overdetermination cases show. We face more or less the opposite situation in the case of the determination idea. It can plausibly be taken to capture a *necessary* condition for causation: under determinism, in order for a set *S* to be the set of *E*'s causes, *S* must nomically determine *E*. But that is obviously not *sufficient* for *S* to be the set of *E*'s causes. (Similarly, under indeterminism it is necessary but not sufficient for a set *S* to be the set of causes of the fact that $\text{ch}_t(E) = p$ that *S* nomically determines that $\text{ch}_t(E) = p$.)

Now, knowledge of a necessary condition for having a certain property often allows us to formulate a sufficient condition as well. Suppose you know that being *N* is a necessary condition for being *A*. And you know that *something* is *A*. Given these assumption, it is a *sufficient* condition for *x* to be *A* that nothing other than *x* is *N*. (If nothing other than *x* is *N*, then nothing other than *x* can be *A*. Hence, given that something is *A*, we can conclude that *x* must be *A*.) Applications of this inference rule are ubiquitous. The detective wants to find out who murdered the victim. Everyone has an unassailable alibi except for the butler, so he must be the culprit. How does this reasoning work? The detective knows that *somebody* must have done it. And she knows a *necessary* condition for someone to be the murderer: having been at the scene of the crime at the time of the murder. If everyone other than the butler fails to meet this *necessary* condition, then that is a *sufficient* condition for the butler to be the murderer. We can call this reasoning procedure the 'method of elimination,' since the detective uses a necessary condition for being the murderer to rule out all possibilities except one.

Necessary conditions can thus give rise to sufficient conditions, thanks to the method of elimination. I claim that nomic determination as a necessary condition for causation gives rise to the (near-)sufficient condition of difference-making by way of the method of elimination. In this section I will restrict my attention to the deterministic case and I will try to give the reader a rough impression of the core idea of this account. It will be the task of sections 4 and 6 to provide a more detailed and rigorous exposition.

I will consider Mill's method of difference, before discussing the other use of the difference-making idea in evaluating causal claim, viz. counterfactual reasoning. Simplifying and idealizing a little, we can describe Mill's method as follows. You observe a scenario (Scenario 1) in which factors *A*, *B*, *C*, and *D* are present at time *t*. At the next instant, *E* obtains. You want to know what caused *E*. Now suppose that you observe another scenario (Scenario 2). In that scenario, *B*, *C* and *D* also obtain, but this time *A* does not obtain. And *E* does not obtain at the next instant. Schematically:

Scenario 1		*Scenario 2*	
t:	$A\ \ B\ \ C\ \ D$	t^*:	$\sim A\ \ B\ \ C\ \ D$
$t+1$:	E	t^*+1:	$\sim E$

Given a certain background assumption to be discussed shortly, these observations support the claim that *A* is a cause of *E* in Scenario 1. A version of the cake example of section 2 can serve as an illustration. You tried to bake a certain kind of cake on two occasions. The first time it tasted chalky and unappealing, but the second time it was delicious. You conclude that you must have done something differently on the two occasions. You look more closely and discover that the first time you used ingredients *A*, *B* and *C*, while the second time you used only *B* and *C*. That was the only difference. You conclude that this difference must be responsible for the difference in flavor. So, your use of ingredient *A* on the first occasion must have been a cause of the chalky taste.

The background assumption you need in order to use the observation of the two scenarios to support your causal conclusion is this: Scenario 2 matches Scenario 1 in all matters of particular fact that are causally relevant to whether *E* occurs, with the possible exception of *A*. There must not be any other causally relevant differences between the two scenarios. (In scientific methodology, we would state this as the requirement that, when we manipulate *A* to determine whether it causes *E*, we need to control for all the other factors that may be causally relevant to *E*.) If the two scenarios differ in other causally relevant ways, then *that* difference might be what is responsible for the fact that *E* obtains in Scenario 1 but not in Scenario 2. Then you cannot put the blame for *E*'s obtaining in Scenario 1 on *A*. In other words, the causal factors with respect to which the two scenarios match each other, i.e. *B*, *C*, and *D*, must include all factors that are causally relevant to *E*, with the possible exception of *A*. That is to say, you have to assume that in Scenario 1 *A*, *B*, *C*, and *D* include all the causes of *E* that obtain at *t*.

I offer the following reconstruction of the method of difference. As we just saw, the method starts from the assumption that

P1. In Scenario 1, the set {*A*, *B*, *C*, *D*} includes all the causes of *E* that obtain at *t*.

While you know that P1 is true, you are not sure whether *all* members of {*A*, *B*, *C*, *D*} are causes of *E* or only some of them. In particular, you do not know whether *A* is a cause of

E. That is what you want to find out. Now in Scenario 2, *B*, *C* and *D* all obtain, but *E* does not obtain at the next instant. That shows that

P2. *B*, *C*, and *D* together do not nomically determine *E*.

But according to (D/d**),

P3. The causes of *E* that obtain at *t* in Scenario 1 jointly nomically determine *E*.

P2 and P3 entail that

C1. In Scenario 1, *B*, *C* and *D* do not include all the causes of *E* that obtain at *t*.

C1 and P1 entail

C2. *A* is a cause of *E* in Scenario 1.

Q.E.D. This argument uses the determination idea to establish the causal conclusion C2 by elimination. (The determination idea, together with Scenario 2, is used to eliminate the possibility that in Scenario 1 *B* – *D* include all the causes of *E* that obtain at *t*. Given P1, that only leaves the possibility that *A* is a cause of *E*.)[26]

In order to use the method of difference to show that *A* is a cause of *E* in Scenario 1, you need to have observed a situation that is just like Scenario 1 in all relevant ways except that *A* does not obtain. But what if you are not lucky enough to have observed

[26] In my exposition of this reconstruction of the method of difference, I have so far ignored the distinction between type causes and token causes mentioned in footnote 19, mainly in order to keep the discussion in the main text simple. Let me briefly describe how the exposition needs to be changed if we pay attention to the distinction.

As mentioned in footnote 19, my discussion is concerned with causation between facts. In my schematic representation of Mill's method, '*A*' – '*E*' stand for type facts, not token facts. In Scenario 1, tokens of *A* – *D* obtain at *t*, and an *E* token obtains at *t*+1. In Scenario 2, tokens of *B* – *D* obtain at *t**, but no *A* token obtains at *t** and no *E* token obtains at *t**+1. Now, we can strengthen principle (5) of footnote 19 in just the way we strengthened (D/d) to obtain (D/d**). That yields the following principle: Let *t* be some time, and assume that a certain token *e* of the type fact *E* obtains at *t*+1. Then,

(5**) If there are tokens of the type facts $X_1, X_2, \ldots, X_n$ that include all the causes of *e* that obtain at *t*, and $y_1, y_2, \ldots, y_n$ are tokens of $X_1, X_2, \ldots, X_n$ that obtain together at another time *t**, then $y_1, y_2, \ldots, y_n$ nomically determine the obtaining of an *E* token at *t**+1.

Premise P1 can be more precisely formulated as the thesis that the tokens of *A* – *D* that obtain in Scenario 1 include all the causes of the *E* token that obtain at *t*. Now, the tokens of *B* – *D* that obtain at *t** in Scenario 2 do not nomically determine that an *E* token obtains at *t**+1. It follows by (5**) that in Scenario 1 the tokens of *B* – *D* do not include all the causes of the *E* token that obtain at *t*. Hence, given the reformulated version of P1, we can conclude that in Scenario 1 the *A* token is one of the causes of the *E* token.

Note that none of the complications described in this footnote arise for counterfactual reasoning (which is the main focus of this paper). As we will see, in counterfactual reasoning we are comparing two scenarios that contain the very same tokens of *B* – *D* (see footnote 27). The notion of a type cause, and the concept of a type fact introduced in footnote 19, are therefore not needed for the discussion of the counterfactual test.

such a scenario? If my reconstruction of the method of difference is right, then that does not really matter. The only function of Scenario 2 is to show that *B*, *C* and *D* by themselves do not nomically determine *E*. But if that is all you want to show, then it is not really necessary to have observed an *actual* scenario that is just like Scenario 1 except that *A* does not obtain. It is sufficient if you can show that in possible worlds where *B*, *C* and *D* obtain without *A* and which conform to the actual laws, *E* does not obtain at the next instant. And you may be able to show this, provided you have enough background knowledge about the laws. Consider the example of the cake again. Suppose that you have never actually tried to make the cake just with ingredients *B* and *C* and with no other ingredients. But suppose that you have used ingredients *B* and *C* in a lot of other ways, and that that has taught you a lot about their nomic roles. In particular, you can conclude from your experiences that in any possible world with the same laws, a cake made with *B* and *C* and no other ingredients does not taste chalky. That is enough to show that in the actual scenario, where you baked the cake with *A*, *B* and *C* and it came out chalky, *B* and *C* by themselves were not nomically sufficient for the chalky taste. You can again use the determination idea to conclude that *B* and *C* cannot include all the causes of the chalky taste that obtain at *t*. Your use of ingredient *A* must have been one of the causes.

In this procedure you are considering a possible situation in which you are not using *A* to make the cake, but which is like the actual situation in all other relevant ways, and which follows the same laws. And you figure out that in such a situation, the cake does not taste chalky. That is a simple version of counterfactual reasoning. Our discussion therefore suggests that the counterfactual test for causal claims is simply an extension of the method of difference. It works in essentially the same way: it relies on the determination idea to establish the causal claim by the method of elimination.[27]

3.2 *The utility of the method of difference and of counterfactual reasoning*

In the previous section, I tried to give the reader a first and rough idea of how the method of difference and the counterfactual test for causal claims work. That idea is enough to allow us to see what makes these tests so useful.

[27] Note an important difference between counterfactual reasoning and the method of difference. In the latter, we are comparing two actual scenarios that contain tokens of the same type causes *B* – *D*, but which contain different token causes (see footnote 26). In counterfactual reasoning, we are comparing an actual and a counterfactual scenario that contain the very same tokens of *B* – *D*.

When you prove that object *X* has property *P* by the method of elimination, you proceed by showing that none of the things other than *X* has *P*. So, even though your ultimate aim is to establish something about *X*, your attention in this process is not on *X* at all. It is on the things *other than X*. You are trying to show that these *other* things do *not* have property *P*. Consider the example of the murder investigation again. The detective shows that the butler must be the killer by showing that everyone else has an unassailable alibi (and therefore cannot have done it). What the detective focuses her attention on in the main part of this procedure is not the actions of the butler at all. Rather, her energies are focused on everyone except for the butler. The detective determines where all of these *other* people were at the time of the crime. Only in the last step of the procedure does she turn her attention back to the butler, to conclude that he must have done it.

I think that this shows something important about what makes the method of elimination so useful. There are many ways of establishing that the butler is the culprit. The most obvious one is to rely on traces you found of the crime that implicate the butler, such as CCTV footage showing him killing the victim, his DNA at the crime scene, the victim's blood on his shoes, and so forth. But what if the butler was extremely careful and did not leave any traces? It is in that case that the method of elimination comes in really handy, since it may allow you to convict the butler nonetheless. All you need is traces of what the *other* people were up to at the time. Suppose you find CCTV footage showing that the gardener was tending to the rhododendrons at the time of the crime, you have reliable witnesses who testify that the general and the professor were reading books in the library, and so forth. That kind of evidence may allow you to show of everyone other than the butler that they are innocent. You can then infer that the butler was guilty. The method of elimination thus allows you to convict the butler in cases where you do not have the kind of evidence required to prove his guilt in a more direct way.

Analogous considerations apply to the method of difference and to the counterfactual test. Suppose that you want to show that ingredient *A* caused your cake to taste chalky. One possible way of doing that is to rely on background knowledge *about A*. Suppose, e.g., that you have used *A* many times before, and you know that, when heated, it always undergoes certain chemical processes that make it taste chalky. This or similar background knowledge about *A* may allow you to conclude that *A* caused the chalky taste. But what if you do not have this background knowledge about *A*?

At that point the method of difference and the counterfactual test prove very useful. When you use either of these procedures to show that *A* caused the chalky taste, the main part of your argument is not concerned with *A* at all. It is concerned with the *other* ingredients. You establish that your use of these *other* ingredients did *not* nomically determine the chalky taste, and that the other ingredients therefore cannot include all the causes of the chalky taste. Only then do you go on to infer that *A* must be one of the causes of the chalky taste. In order to apply this method, you do not need to know much about *A* at all. All you need is sufficient knowledge about the *other* ingredients to know that they did not nomically determine the chalky taste. Maybe you know that much because you made another cake *just* with *B* and *C* and it did not taste chalky. Or maybe you know enough about the chemical composition of *B* and *C* and about the laws of chemistry to know that cakes that are made *just* with *B* and *C* do not taste chalky. The important point is: if you have such knowledge of *B* and *C*, then that is enough to establish that *A* caused the chalky taste. You need to know next to nothing about *A* itself.

4. The counterfactual test under determinism

The simple counterfactual test considered in section 3.1 works only under determinism, and only if you already know a lot about which of the matters of particular fact that obtain at *t* are causes of *E* (more precisely: you need to know that *A* – *D* include all of *E*'s causes that obtain at *t*). And it rests on principle (D/d**). I will argue in this section that, by modifying the rules for the test somewhat, we can obtain a new test that does not require much background knowledge about the causes of *E*, and which only uses a weaker version of the determination idea, viz. (D/d*). Moreover, in section 6, we will see that there is a version of this generalized test that can be used under indeterminism. As it turns out, the modified test is just the counterfactual test we use in ordinary life (governed by the rules described in sections 1.1 and 1.3).

Suppose that determinism is true. And assume that at *t*, Fred insults his boss. A week later he gets fired. We want to know whether the insult caused his dismissal.[28] Let

[28] Expressions like 'the insult' and 'the dismissal' can be used either to refer to events or to refer to facts (such as the fact that Fred insulted his boss at time *t*, or that Fred got dismissed at time *t**). Since in this paper I am concerned with token causes that are facts (see footnote 19), I am using the terms in the second way.

$S_{\leq t}$ = the set containing all matters of particular fact[29] that obtain at times up to (and including) t,

$S_{\leq t}^{-}$ = (pronounced '$S_{\leq t}$ minus') the set containing all members of $S_{\leq t}$ except for Fred's insult,

L = the set of all laws of nature,

$LS_{\leq t}^{-}$ = the union of $S_{\leq t}^{-}$ and L.

(D/d*) tells us that those members of $S_{\leq t}$ that are causes of Fred's dismissal jointly nomically determine the dismissal. Now suppose that we were able to show that

(1) $LS_{\leq t}^{-}$ does not determine Fred's dismissal,[30]

i.e. that the members of $S_{\leq t}^{-}$ do not nomically determine Fred's dismissal. Then we would be able to conclude that some of the causes of the dismissal that are in $S_{\leq t}$ are not in $S_{\leq t}^{-}$. Since the insult is the only member of $S_{\leq t}$ that is not in $S_{\leq t}^{-}$, it would follow that the insult is a cause of the dismissal.

Unfortunately, this strategy for supporting the causal claim requires some serious revision, for (1) is not true. $S_{\leq t}^{-}$ includes all the matters of particular fact obtaining before t (i.e., before the time of the insult). And under determinism the history of the world before t contains factors that are nomically sufficient for Fred's dismissal, whether or not the insult is one of the causes of the dismissal. If the insult *is* a cause of the dismissal, then $S_{\leq t}^{-}$ might contain something like the following factors:

the fact that at 11:59:59 Fred forms the intention to utter the insult,

the absence of factors that could prevent Fred from carrying out his intention,

the fact that Fred's boss is very unforgiving,

the fact that Fred's boss has a very low threshold for firing employees,

and so forth. Complete this list the right way and you obtain a set of factors that is nomically sufficient for Fred's insult and for his boss's being disposed to fire Fred if Fred insults him. Taken together, these factors are nomically sufficient for Fred's dismissal. Suppose next that the insult is *not* one of the causes of the dismissal, but that the

[29] By 'matters of particular fact' I mean, very roughly speaking, the facts about what kinds of goings-on fill space-time.

[30] When I say in what follows that a certain set *S determines E*, I will mean that *S* necessitates *E*.

dismissal was instead caused by Fred's poor performance during the previous month. Then $S_{\leq t}^{-}$ may contain something like the following factors:

the fact that Fred's boss remembers Fred's poor performance,

the fact that the boss thinks that his company must save money,

the fact that the boss thinks that the best way to save money is to lay off some employees,

and so on. These factors nomically determine that Fred gets fired. There is, however, a conspicuous difference between the two cases. If the insult is a cause of the dismissal, then the factors in $S_{\leq t}^{-}$ that nomically determine the dismissal do so *by way of* nomically determining the insult. (There is a causal chain from the factors in $S_{\leq t}^{-}$ to the dismissal, and this causal chain runs through the insult.) Matters are different if the insult is *not* a cause of the dismissal. Then the factors in $S_{\leq t}^{-}$ still nomically determine the dismissal, but not by way of nomically determining the insult. (There is still a causal chain from the factors in $S_{\leq t}^{-}$ to the dismissal, but the causal chain does not run through the insult.) The question, then, is simply whether the factors in $S_{\leq t}^{-}$ nomically determine Fred's dismissal by way of nomically determining the insult, or whether they nomically determine his dismissal in some other way. If we can rule out the second possibility, then that shows that the insult is a cause of the dismissal. So, what needs to be shown in order to support the causal claim is that

(2) $LS_{\leq t}^{-}$ does not determine Fred's dismissal *in any way other than by determining the insult*.

But how can we show that (2) is true? One strategy is to consider the set $LS_{\leq t}^{-}*$ that we obtain by weakening $LS_{\leq t}^{-}$ *just* enough to ensure that the resulting set does not determine the insult (while otherwise leaving the set unchanged). If $LS_{\leq t}^{-}$ determines the dismissal only by way of determining the insult, then we should expect that $LS_{\leq t}^{-}*$ does *not* determine the dismissal. (For $LS_{\leq t}^{-}*$ does not determine the insult.) On the other hand, if there are factors in $LS_{\leq t}^{-}$ that determine the dismissal other than by way of determining the insult, then (since almost all members of $LS_{\leq t}^{-}$ are also in $LS_{\leq t}^{-}*$) these factors are likely also in $LS_{\leq t}^{-}*$, so that $LS_{\leq t}^{-}*$ determines the dismissal. Hence, one way of trying to find out whether $LS_{\leq t}^{-}$ determines the dismissal other than by way of determining the insult is to find out whether $LS_{\leq t}^{-}$* determines the dismissal.

It would be nice if we could find some general rules for constructing the set $LS_{\leq t}^{-}$*. The task can be thought of in terms of possible worlds. Consider possible worlds where Fred does not insult his boss ('no-insult worlds'), but which otherwise match our world as closely as possible with respect to the factors in $LS_{\leq t}^{-}$. The set $LS_{\leq t}^{-}$* can then be defined as the set of all elements of $LS_{\leq t}^{-}$ that obtain in such a world. And we can show that $LS_{\leq t}^{-}$* does not determine the dismissal by showing that Fred does not get dismissed in such a world.

What we need, then, is a general recipe for constructing such a world. Now, remember that $LS_{\leq t}^{-}$ contains two kinds of fact: the laws of nature, and the matters of particular fact up to *t* (except for the insult). In consequence, the no-insult worlds we are looking for are those that satisfy two desiderata: *ceteris paribus*, they ought to conform as closely as possible to the actual laws; and, *ceteris paribus*, they ought to match our world as closely as possible up to *t*. These, of course, are just the two criteria for closeness that we use in ordinary counterfactual reasoning under determinism (as described in section 1.1). We already know what the no-insult worlds look like that provide the best trade-off between the two desiderata: they are like the actual world until shortly before *t*, then diverge from our world by a small miracle that ensures that Fred does not utter the insult, and afterwards unfold in accordance with the actual laws. If Fred does not get dismissed in these worlds, then we can infer that the insult is a cause of the dismissal.

This inference is fallible. We take the fact that Fred does not get fired in the closest no-insult worlds to show that $LS_{\leq t}^{-}$ does not determine the dismissal other than by way of determining the insult (and hence that the insult is a cause of the dismissal). But, while the first proposition provides support for the second, there are cases where the first is true while the second is false. For not *all* members of $LS_{\leq t}^{-}$ obtain in the closest no-insult worlds. So, it could be that the insult is not a cause of the dismissal, and that there are therefore certain factors in $LS_{\leq t}^{-}$ that determine the dismissal other than by way of determining the insult, but that these factors are among the few members of $LS_{\leq t}^{-}$ that do not obtain in the closest no-insult worlds (i.e., that they are among the members of $LS_{\leq t}^{-}$ that were removed from $LS_{\leq t}^{-}$ in the course of forming $LS_{\leq t}^{-}$*). Then it may be that Fred does not get fired in the closest no-insult worlds. That is just what happens in the backtracking cases considered in section 1.1.

Consider an example of this problem. Suppose that in the actual world the utterance of the insult is caused by the firing of certain neurons. In the closest no-insult worlds, one of these neurons miraculously fails to fire. (Thus, the neuron firing is one of the factors that we remove from $LS_{\leq t}^{-}$ in the course of forming $LS_{\leq t}^{-}*$.) Now suppose that in our world the firing of the same neuron also caused Fred's hand to twitch so that he inadvertently spilled some of his drink on his boss's new suit. Distracted by this incident, the boss was not listening when Fred uttered his insult, and therefore did not know that he had been insulted. But a week later, when he remembered how Fred ruined his suit, he got so enraged that he decided to fire Fred. In the closest no-insult worlds, the insult neuron does not fire. Fred's hand does not twitch and he does not spill the drink. His boss does not get mad, and Fred keeps his job. The dismissal counterfactually depends on the insult, even though it was not caused by it.

Although the counterfactual test is defeasible for the reasons described, there is little risk in practice that it will lead us to accept false conclusions about causation. As the example of the last paragraph illustrates, in cases where the counterfactual dependence holds for backtracking reasons, we typically have no way of establishing the relevant counterfactual except by *backtracking reasoning*, i.e. reasoning from the antecedent back in time to conclusions about what the pre-antecedent history must have been like in order for the antecedent to become true, and then forward from there to the post-antecedent time. But if that is how we establish the counterfactual, then it should be clear to us that we cannot infer that the insult is a cause of the dismissal.

5. Over-determination and preemption revisited

Counterfactual dependence between distinct matters of particular fact is a near-sufficient condition for causation, but it is not a necessary condition, as preemption and over-determination cases show. Hence, where *C* and *E* are distinct matters of particular fact, we can gain strong evidence that *C* caused *E* if we can show that *E* counterfactually depends on *C*, but we cannot show that *C* did *not* cause *E* by showing that *E* is counterfactually *in*dependent of *C*. The counterfactual test for causation works only in one direction.

That is not at all surprising on the assumption that the test is an application of the method of elimination, for it is a characteristic feature of that method that it often yields only a one-way test. The method, to recall, starts from the premise that having a certain property N is a necessary condition for being A. Add to this the assumption that something is A, and you obtain a *sufficient* condition for something to be A: the condition that nothing else is N (call this condition 'S'). But note that S need not be a *necessary* condition for being A. That is to say, if x does not meet condition S—i.e., if something other than x *is* N—it does not follow that x is not A. All that follows is that the relevant other thing meets one specific necessary condition, viz. N, for being A. But, unless N is not only a necessary but also a sufficient condition for being A, it does not follow that the other thing is A, let alone that x is not A. The detective example of section 3.1 illustrates this. Having been at the scene of the crime at the time when the victim was killed is a necessary, but not a sufficient, condition for someone to be the culprit. So, if we can show that no one other than the butler meets this necessary condition, then that is a sufficient condition for the butler to be guilty. However, it is not a *necessary* condition for the butler to be guilty that no one else was at the scene of the crime at the time of the murder. After all, the butler could have murdered the victim in the presence of someone else (an accomplice or an innocent bystander).

My account explains over-determination and preemption cases as instances of the same phenomenon. Let A and E be distinct matters of particular fact, with A obtaining at t and E obtaining at some later time, and let $S_{\leq t}^{-}$ be the set of all matters of particular fact up to (and including) t except for A. It is a necessary (but not sufficient) condition for $S_{\leq t}^{-}$ to contain all the causes of E up to t that $S_{\leq t}^{-}$ nomically determines E other than by way of nomically determining A. If we can show that $\sim E$ in the closest $\sim A$-worlds, then we have strong reasons for thinking that $S_{\leq t}^{-}$ does not meet this necessary condition for containing all the causes of E up to t, and that A must therefore be one of these causes. But if E is true in the closest $\sim A$-worlds, then the most we can conclude is that $S_{\leq t}^{-}$ *does* meet our necessary condition for including all the causes of E that obtain up to t. But since that condition is not a *sufficient* condition, it does *not* follow that $S_{\leq t}^{-}$ contains all of E's causes up to t. A may still be one of these causes.

That is just what happens in cases of over-determination and preemption. First, over-determination. Fred's and Susie's bricks simultaneous collide with the bottle, each causing sufficient damage to shatter the bottle. Consider the set $S_{\leq t}^{-}$ that contains all matters of particular fact up to the time of Susie's throw, except for that throw itself. Since Susie's throw is one of the causes of the shattering and $S_{\leq t}^{-}$ does not contain Susie's throw, $S_{\leq t}^{-}$ does not contain all the causes up to the time of Susie's throw. But $S_{\leq t}^{-}$ contains Fred's throw and certain background factors that nomically ensure that his throw will be followed by the shattering of the bottle. These factors nomically determine that the bottle will shatter; and they do so in a way other than by nomically determining that Susie throws her brick. Hence, $S_{\leq t}^{-}$ meets our necessary condition for including all the causes of the shattering up to the time of Susie's throw. But since this necessary condition is not also a sufficient condition, we cannot conclude that $S_{\leq t}^{-}$ contains all the causes of the shattering up to the time of Susie's throw, and therefore cannot infer that Susie's throw was not a cause of the shattering.

Similarly in preemption cases. Suppose that Susie throws her brick first and shatters the bottle. Fred sees this, decides that the job has already been done, and does not throw his brick. Susie's throw is a preempting cause of the shattering. Let '$S_{\leq t}^{-}$' stand for the same set as before. Again, $S_{\leq t}^{-}$ does not contain all the causes of the bottle's shattering up to the time of Susie's throw, since it does not contain Susie's throw. But $S_{\leq t}^{-}$ contains Fred's intention to break the bottle, and background factors that nomically determine that nothing will prevent him from carrying out this intention except something else's shattering the bottle first. These factors nomically determine that the bottle will shatter; and they do so in a way other than by nomically determining that Susie will throw her brick. Hence, $S_{\leq t}^{-}$ meets our necessary condition for including all the causes of the shattering up to the time of Susie's throw. But, once again, since this necessary condition is not also a sufficient condition, we cannot conclude that $S_{\leq t}^{-}$ contains all the causes of the shattering up to the time of Susie's throw, and therefore cannot infer that Susie's throw was not a cause of the shattering.

6. The counterfactual test under indeterminism

Next, an indeterministic version of the counterfactual test. Fred insults his boss on Monday. On Tuesday morning, his boss learns that some of the employees have to be laid off to save money. He looks over the list of employees, trying to decide whom to fire. He ends up picking Fred. Suppose that a moment before the decision (at time *t*), it was not yet settled what decision he would make. But, given the boss's grudge against Fred, it was already very likely that Fred would be the one to be laid off. Let us say that the chance was *p*. We are wondering whether this probability is due to the insult.

Let

$S_{\leq t}$ = the set of all matters of particular fact obtaining no later than *t*,

$S_{\leq t}^{-}$ = the set of all members of $S_{\leq t}$ except for the insult,

L = the set of all laws of nature,

$LS_{\leq t}^{-}$ = the union of $S_{\leq t}^{-}$ and L.

Barring backwards causation, the factors in $S_{\leq t}$ include all the causes of the fact that $\text{ch}_t(\text{dismissal}) = p$.[31] Now suppose that we were able to show that

(3) $LS_{\leq t}^{-}$ does not determine that $\text{ch}_t(\text{dismissal}) = p$.

Then we could apply (D/i) to conclude that $S_{\leq t}^{-}$ does not include all the causes of the fact that $\text{ch}_t(\text{dismissal}) = p$. It would follow that there are some causes of the fact that $\text{ch}_t(\text{dismissal}) = p$ that are in $S_{\leq t}$ but not in $S_{\leq t}^{-}$. Hence, the insult must be one of the causes of the fact that $\text{ch}_t(\text{dismissal}) = p$.

Unfortunately, this strategy for supporting the causal claim requires some serious revision, for (3) is not true. Note that $S_{\leq t}^{-}$ contains all matters of particular fact that obtain between the time of the insult and *t*. And these include some nomically sufficient causes of the fact that $\text{ch}_t(\text{dismissal}) = p$, whether or not there is a causal connection between the insult and the fact that $\text{ch}_t(\text{dismissal}) = p$. Suppose first that there *is* a causal connection. Then $S_{\leq t}^{-}$ contains such factors as the boss's memories of the insult, his anger at the incident, and so forth. These factors nomically determine that $\text{ch}_t(\text{dismissal}) = p$. Next, suppose that there is *no* causal connection between the insult and the fact that $\text{ch}_t(\text{dismissal}) = p$, but that the latter fact was due to Fred's poor performance. Then $S_{\leq t}^{-}$ contains such factors as the boss's memories of Fred's poor performance, his eagerness to

[31] Note that I am not assuming that the matters of particular fact *before t* include all the causes of the fact that $\text{ch}_t(\text{dismissal}) = p$, or even that they nomically determine that $\text{ch}_t(\text{dismissal}) = p$. I think that some of the causes of facts about the chances at *t* may obtain *at t*. Suppose that at *t* a certain particle has a certain chance of decaying within the next year. This chance is partly determined by the fact that the particle is in a certain state *at t*. Moreover, the history *before t* may not nomically determine that the particle is in that specific state at *t* (the particle may only have entered into that state at *t*, and it may have been a matter of chance that it did so). So, the world's history before *t* does not include all the causes of the particle's chance at *t* of decaying within a year, and does not nomically determine that chance.

save money, his belief that the best way of doing so is to lay off some employees, and so forth. Again, these factors nomically determine that $\mathrm{ch}_t(\text{dismissal}) = p$. There is, of course, a crucial difference between the two cases. In the first case, the boss has memories of the insult and is angry only *because* Fred insulted him. The factors between the time of the insult and t that nomically determine that $\mathrm{ch}_t(\text{dismissal}) = p$ include some matters of particular fact that themselves causally depend on the insult. In the second case, by contrast, the factors between the insult and t that nomically determine that $\mathrm{ch}_t(\text{dismissal}) = p$ are causally independent of the insult.

The crucial question that we need to answer to determine whether the insult figures in the causal history of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$, therefore, is whether those factors in $S_{\leq t}^{-}$ that are causally *in*dependent of the insult are nomically sufficient for $\mathrm{ch}_t(\text{dismissal}) = p$. More formally, let $S_{\leq t}^{\text{insult}}$ be the set of those matters of particular fact in $S_{\leq t}^{-}$ that are causally independent of the insult, and let $LS_{\leq t}^{\text{insult}}$ be the union of $S_{\leq t}^{\text{insult}}$ and the set of all laws. Suppose that we can show that

(4) $LS_{\leq t}^{\text{insult}}$ does not determine that $\mathrm{ch}_t(\text{dismissal}) = p$.

Then we can use (D/i) to conclude that $LS_{\leq t}^{\text{insult}}$ does not include all the causes of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$. Since $LS_{\leq t}$ *does* include all the causes, it follows that the causes include some of the elements of $LS_{\leq t}$ that are not in $LS_{\leq t}^{\text{insult}}$. Hence, either the insult, or some matters of particular fact that causally depend on the insult, are among the causes of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$. In either case, it follows that the insult figures in the causal history of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$.[32]

[32] By 'C figures in the causal history of E,' I mean that C stands in the ancestral relation of causation to E (i.e., that E causally depends on C, in the sense defined in footnote 14). What my discussion shows, then, is that (4) entails

(6) The insult stands in the ancestral relation of causation to the fact that $\mathrm{ch}_t(\text{dismissal}) = p$.

Whether or not that by itself is enough to conclude that

(7) The insult is a cause of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$

depends on whether causation is a transitive relation. That is a controversial issue, as mentioned in footnote 14. Without taking a stand on this issue, therefore, I cannot claim that my discussion of the counterfactual test under indeterminism shows more than this: the counterfactual dependence of the fact that $\mathrm{ch}_t(\text{dismissal}) = p$ on the insult supports the assumption that (6) holds. I think, however, that even if causation is not transitive, we can still use the counterfactual test to support the stronger claim (7). But a little more work is required to show this. What we need to do is to consider the different kinds of counterexample to transitivity (i.e. those cases where (6) holds but (7) does not) one by one, and show that in these cases the fact that $\mathrm{ch}_t(\text{dismissal}) = p$ does not counterfactually depend on the insult. We can then infer that all cases

There is one additional complication. Indeterminism is the thesis that not everything is pre-determined. It is not the thesis that *nothing* is pre-determined. Some matters of particular fact may still be determined by earlier matters and the laws. Now suppose that the insult figures in the causal history of the fact that ch_t(dismissal) = p. Suppose further that Fred's insult was nomically pre-determined by factors obtaining before the insult, and that the same is true of all the other factors that obtain between the insult and t and which causally contributed to the fact that ch_t(dismissal) = p. Then $LS_{\leq t}^{\text{insult}}$ determines that ch_t(dismissal) = p. I.e., (4) is false. So, we cannot use the strategy described in the last paragraph to show that the insult figures in the causal history of the fact that ch_t(dismissal) = p.

But we can modify the strategy so that it applies even to cases like the one just considered. We can use the same trick as in the deterministic case. We weaken $LS_{\leq t}^{\text{insult}}$ just enough to ensure that the resulting set $LS_{\leq t}^{\text{insult}*}$ does not determine the insult. If we can show that $LS_{\leq t}^{\text{insult}*}$ does not determine that ch_t(dismissal) = p, then we can conclude that the factors in $LS_{\leq t}^{\text{insult}}$ do not determine that ch_t(dismissal) = p in any way other than by determining the insult. Then the insult must figure in the causal history of the fact that ch_t(dismissal) = p.

We can give a unified description that covers both the version of the reasoning strategy that proceeds by showing that (4) is true and the more complex variant discussed in the last paragraph. Consider the no-insult worlds that match our world most closely with respect to the members of $LS_{\leq t}^{\text{insult}}$. If $LS_{\leq t}^{\text{insult}}$ does not determine the insult, then all factors in $LS_{\leq t}^{\text{insult}}$ obtain in these worlds. And if we can show that ch_t(dismissal) $\neq p$ in these worlds, then it follows that (4) is true, and that the insult must therefore figure in the causal history of the fact that ch_t(dismissal) = p. If $LS_{\leq t}^{\text{insult}}$ *does* determine the insult, then not all factors in $LS_{\leq t}^{\text{insult}}$ obtain in the worlds under consideration. Then we can define $LS_{\leq t}^{\text{insult}*}$ as the set of those members of $LS_{\leq t}^{\text{insult}}$ that obtain in these worlds. And if we can show that ch_t(dismissal) $\neq p$ in these worlds, then it follows that $LS_{\leq t}^{\text{insult}*}$ does not determine that ch_t(dismissal) = p. That gives us strong reasons for thinking that $LS_{\leq t}^{\text{insult}}$ does not determine the fact that ch_t(dismissal) = p other than by way of

of counterfactual dependence where (6) is true are also cases where (7) holds. Hence, if the fact that ch_t(dismissal) = p counterfactually depends on A, then that supports (7) as strongly as it supports (6).

This additional work can be done in a straightforward and fairly mechanical way, but it requires a somewhat lengthy discussion of a large number of different examples. I omitted it for reasons of space.

determining the insult, and therefore supports the conclusion that the insult figures in the causal history of the fact that ch$_t$(dismissal) = p.

So, what we need to do to support that causal claim is to show that ch$_t$(dismissal) $\neq p$ in the no-insult worlds that match our world most closely with respect to the factors in $LS_{\leq t}^{\text{insult}}$. Now $LS_{\leq t}^{\text{insult}}$ contains (i) the laws, as well as (ii) the matters of particular fact that obtain up to t except for the insult and matters that causally depend on the insult. (ii), in turn, includes all matters obtaining up to the time of the insult except for the insult itself, as well as those between the insult and t that are causally independent of the insult. Hence, the no-insult worlds that match our world most closely with respect to the facts in $LS_{\leq t}^{\text{insult}}$ are those that offer the best trade-off between conformity to the laws, match up to the time of the insult and match between the insult and t in matters that are causally independent of the insult. This description fits the no-insult worlds that are closest by the standards governing ordinary counterfactual reasoning under indeterminism (as described in section 1.3). What we need to do, then, is to show that ch$_t$(dismissal) $\neq p$ in the no-insult worlds that are closest by these standards. Having shown that, we can conclude that the insult figures in the causal history of the fact that ch$_t$(dismissal) = p.

By studying how the chance of E at t counterfactually depends on C, we can establish claims about the causal connection between C and E's chance at t. But as we saw in section 1.2, we cannot in the same way establish that C figures in the causal history of E. That can easily be explained on my account. Any use of the counterfactual test must rest on a suitable version of the determination idea. In order to infer from the fact that E's chance at t counterfactually depends on C that C figures in the causal history *of the fact that ch$_t$(E) = p*, we need thesis (D/i). By contrast, in order to infer from the same counterfactual that C figures in the causal history *of E*, we would need a different version of the determination idea, namely thesis ~~(D/i)~~ of section 2. Given ~~(D/i)~~, the inference would be straightforward. Let $S_{\leq t}^{C}$ be the set of all matters of particular fact obtaining up to t except for C and the factors that causally depend on C. If we can show that the chance of E at t counterfactually depends on C, then that supports the claim that the factors in $S_{\leq t}^{C}$ do not nomically determine that ch$_t$(E) = p, or at least do not nomically determine that ch$_t$(E) = p other than by way of nomically determining C. Given ~~(D/i)~~, we would be able to conclude that $S_{\leq t}^{C}$ does not contain all the causes of E, so that either C or some factors that causally depend on C must be among E's causes. In either case it would follow that *C* figures in the causal history of *E*. Without ~~(D/i)~~, however, this inference does not go through. And, as we saw in section 2, ~~(D/i)~~ is false. We can appeal to the falsity of ~~(D/i)~~ to *explain* why the inference fails.

Of course, even the counterfactual test for claims about the causal history of chances is subject to the usual limitations in cases of over-determination, preemption, and backtracking. These can be explained in the same way as under determinism.

7. Conclusion

Counterfactuals frequently serve as our guides to the causal facts. Counterfactual analyses of causation give one explanation of this fact. My account presents another. Let me use the results of this paper to compare the two approaches. Over-determination and preemption cases show that counterfactual dependence between distinct matters of particular fact is not a necessary condition for causation, and backtracking cases show that it is not a sufficient condition. That presents a serious problem for the counterfactual analysis of causation. The viability of that analysis depends on whether it is possible to find some more complex pattern of counterfactual dependencies that is necessary and sufficient for causation. But none of the examples present a difficulty for my position, according to which counterfactual dependence merely provides evidence for causal connections, but does not constitute them. All we need to conclude from over-determination and preemption cases is that counterfactual dependence is only a one-way test for causal connections, not a two-way test: it can establish, but not refute, causal claims. And backtracking cases merely show that it is not quite an *infallible* method for *establishing* the existence of causal connections either. But it is still very reliable and, as we saw in section 4, extremely unlikely to yield false causal conclusions in ordinary cases. What is more, my account predicts and explains all the limitations of the counterfactual test. Hence, far from presenting a difficulty for my position, these limitations confirm the account. The indeterministic case, as we saw in section 1.2, presents additional difficulties for counterfactual analyses, but, once again, the relevant phenomena are just what we would expect on my account (as shown at the end of section 6).

The need for causal notions in the theory of counterfactuals threatens the counterfactual analysis of causation with circularity. But it presents no difficulty for my view that counterfactual reasoning serves as a test for causal claims. All we need to conclude is that counterfactual thinking cannot in general create causal knowledge from scratch. Given that counterfactuals have causal truth-conditions, causal knowledge is often required to establish a counterfactual. In such cases, we cannot acquire causal knowledge by counterfactual reasoning unless we already have some causal knowledge to begin with. But there is no regress. For the causal knowledge required for our counterfactual reasoning is different from that which we acquire as a result of it. We use one item of causal knowledge to gain another. In that way, counterfactual reasoning extends our stock of causal knowledge. And that is what makes it useful. Moreover, not only is my account *compatible* with the fact that counterfactual thinking often relies on prior causal knowledge, but my discussion in section 6 *explains why* causal notions enter into the rules of counterfactual reasoning in the way they do.

My view provides a unified account of the workings of the counterfactual test and of Mill's method of difference (and hence of the method of controlled experiments). It also gives us a starting point for a functional explanation of why we have in our conceptual repertoire a counterfactual connective that is governed by the specific standards of closeness of worlds described in sections 1.1 and 1.3. For in sections 4 and 6 I discussed what set of rules should guide us if we want to use the determination idea to establish a causal claim by the method of elimination. It turned out that the rules that are best for this purpose are just those that govern our actual practice of counterfactual reasoning. That is to say, counterfactual reasoning proceeds in just the way we would expect on the assumption that it developed as a method of supporting causal claims that uses the determination idea and the method of elimination. And that supports the claim that the practice did develop, at least in part, for this purpose. That is not to deny that counterfactual reasoning serves other functions as well (e.g. in decision making). Counterfactual reasoning may have been molded under the influence of a variety of functional pressures that converged on the same set of rules. In that case, the practice emerged for more than one reason. One of them is given in this paper.[33]

References

[33] I am grateful to Adam Elga, David Chalmers, Ned Hall, Mark Johnston, James Joyce, Igal Kvart, David Manley, Laurie Paul, Peter Railton, Gideon Rosen, Jonathan Schaffer, and Fritz Warfield for comments and suggestions concerning the ideas presented in this paper. I am also indebted to the audiences of talks on the material that I gave at the University of Michigan in Ann Arbor, Notre Dame, MIT, Victoria University in Wellington, Australian National University, Princeton, and at the Arizona Ontology Conference 2009, to the participants of two graduate seminars that I gave at Michigan, and to an anonymous referee for *Noûs*.

Adams, Ernest. (1975) *The Logic of Conditionals*. Dordrecht: Reidel.

Bennett, Jonathan. (1984) “Counterfactuals and Temporal Direction.” *Philosophical Review* 93: 57–91.

——. (1988) *Events and their Names*. Indianapolis: Hackett Publishers.

——. (2003) *A Philosophical Guide to Conditionals*. Oxford: Clarendon.

Campbell, Joseph K., Michael O’Rourke and Harry S. Silverstein. (eds.) (2007) *Causation and Explanation*. Cambridge, MA: MIT Press

Collins, John, Ned Hall and Laurie Paul. (eds.) (2004) *Causation and Counterfactuals*. Cambridge, MA: Bradford Book / MIT Press.

Davidson, Donald. (1980) “Causal Relations,” orig. 1967, in *Essays on Actions and Events*. Oxford: Clarendon Press, pp. 149-62.

Dowe, Phil and Paul Noordhof. (eds.) (2003) *Causation and Counterfactuals*. London: Routledge.

Edgington, Dorothy. (1995), “On Conditionals,” *Mind* 104, pp. 235–329.

——. (2003) “Counterfactuals and the Benefit of Hindsight,” in Dowe and Noordhof (2003), 12-27.

Hall, Ned. (2004a) “Two Concepts of Causation.” in Collins et al. 2004, 225–277.

——. (2004b) “Causation and the Price of Transitivity.” in Collins et al. (2004), 181–205.

——. (2007) “Structural Equations and Causation.” *Philosophical Studies* 132: 109-136.

Halpern, Joseph Y. and Judea Pearl (2005). “Causes and Explanations: A Structural-Model Approach. Part 1: Causes.” *British Journal for the Philosophy of Science* 56: 843-887.

Hiddleston, Eric. (2005) “A Causal Theory of Counterfactuals.” *Nous* 39: 632-657.

Hitchcock, Christopher. (2001) “The Intransitivity of Causation Revealed in Equations and Graphs.” *Journal of Philosophy* 98: 273-299.

——. (2004) “Do All and Only Causes Raise the Probabilities of Effects?”, in Collins et al. 2004, 403–417.

Hume, David. (1995) *An Inquiry Concerning Human Understanding*. Upper Saddle River: Prentice Hall.

Jackson, Frank. (1977) “A Causal Theory of Counterfactuals.” *Australasian Journal of Philosophy* 55: 3–21.

Kim, Jaegwon. (1973) “Causation, Nomic Subsumption, and the Concept of Event.” *Journal of Philosophy* 70: 217-36.

Kment, Boris. (2006a) “Counterfactuals and Explanation.” *Mind* 115: 261-310.

Kment, Boris. (2006b) “Counterfactuals and the Analysis of Necessity.” *Philosophical Perspectives* 20: 237-302.

Lewis, David. (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.

——. (1986a) *Philosophical Papers*. New York / Oxford: Oxford University Press, vol.ii.

——. (1986b) “Causation,” in Lewis (1986a), 159–213.

——. (1986c) “Counterfactual Dependence and Time’s Arrow,” in Lewis (1986a), 32–66.

——. (1986d) “Events,” in Lewis (1986a), 241-69.

——. (2004) “Causation as Influence,” in Collins et al. (2004), 75–107.

Mackie, John L. (1974) *The Cement of the Universe*. Oxford: Clarendon.

Mårtensson, Johan. (1999) *Subjunctive Conditionals and Time*. Acta Universitatis Gothoburgiensis.

McDermott, Michael. (1995) “Redundant Causation.” *British Journal for the Philosophy of Science* 46: 523–544.

Mellor, David H. (1995) *The Facts of Causation*. London: Routledge Press.

——. (2004), “For Facts as Causes and Effects,” in Collins et al. (2004), 309-323.

Menzies, Peter. (1989a) “Probabilistic Causation and Causal Processes: A Critique of Lewis.” *Philosophy of Science* 56: 642–663.

——. (1989b) “A Unified Account of Causal Relata,” *Australasian Journal of Philosophy* 67, pp. 59-83

Mill, John Stuart. (1956) *A System of Logic, Ratiocinative and Inductive*. London/New York: Longmans, Green.

Nolan, Daniel. (1997) “Impossible Worlds: A Modest Approach.” *Notre Dame Journal of Formal Logic* 38: 535-73.

Paul, Laurie. (2004), “Aspect Causation,” in Collins et al. 2004, 205-24.

Pearl, Judea. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge, UK/New York: Cambridge UP.

Ramachandran, Murali. (1997) “A Counterfactual Analysis of Causation.” *Mind* 151: 263–277.

Schaffer, Jonathan. (2000) “Overlappings: Probability-Raising Without Causation.” *Australasian Journal of Philosophy* 78: 40-46.

——. (2004a) “Trumping Preemption,” in Collins et al. (2004), 59–75.

——. (2004b), “Counterfactuals, Causal Independence and Conceptual Circularity,” *Analysis* 64, pp. 299-309

Stalnaker, Robert. (1968) “A Theory of Conditionals.” *Studies in Logical Theory, American Philosophical Quarterly Monograph Series* 2. Oxford: Blackwell, 98–112.

Strevens, Michael. (2007) “Mackie Remixed,” in Campbell et al. (2007), 93–118.

Tichý, Pavel. (1976) “A counterexample to the Stalnaker-Lewis analysis of counterfactuals.” *Philosophical Studies* 29: 271–273.

Wasserman, Ryan. (2006) “The Future Similarity Objection Revisited.” *Synthese* 150: 57-67.

Woodward, James. (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford UP.

Yablo, Stephen. (2004) "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation," in Collins et al. 2004