OXFORD PHILOSOPHICAL SOCIETY – MEMBERS' WEEKEND, AUGUST 2018

SCRIPT FOR THE PRESENTATION

TITLE: Counterfactual Theories of Causation ("CTCs")

PRESENTER: Michael Donnan

[**Slide 1**: Title, Presenter, Occasion and Date]

Introduction

If you are expecting me to start by discussing David Hume's second definition of causation, I regret I must disappoint you. For I am going to start instead with a little tale of a minor mishap. **[Slide 2]**

Imagine, if you will, that you are performing experiments in a chemistry laboratory. You have your Bunsen burner properly regulated to give a pale blue flame. **[Same slide 2, click]** You dip your spatula into a jar of strontium carbonate with a view to transferring some of the crystals to a beaker. **[Same slide 2, click]** But as you attempt the transfer, you sneeze **[Same slide 2, click]** and the force of this sneeze blows the crystals off the spatula and into the flame **[Same slide 2, click]**, which thereupon turns a pleasing shade of red **[Same slide 2, click].**

I want to isolate a sequence of three distinct events: first, the sneeze: second, the blowing of the crystals into the flame; and, third, the flame's colour change. Let's consider the second and third events. Each constitutes a difference made to the immediately pre-existing state of affairs and it seems natural to accept that what made the difference – the difference-maker - was the preceding event. So, the flame changed colour owing to the crystals being blown into it, and the crystals were blown into the flame owing to the sneeze. But it also seems natural to take it that without the crystals being blown into the flame, the flame would not have changed colour

and without the sneeze the crystals would not have been blown into the flame. In other words, *if* the difference-making event – the difference-maker - had *not* occurred, then the difference would *not* have occurred.

We seem to have picked out a dependence relation. Let's generalise from this and formulate an analysis of this kind of dependence. **[Slide 3]**

> Where ***c*** and ***e*** are distinct events, ***e*** somewise depends on ***c*** if and only if (1) if it were the case that ***c*** occurs, then it would be the case that ***e*** occurs, and (2) if it were the case that ***c*** does not occur, then it would be the case that ***e*** does not occur.

At this point, I need to introduce the notion of a counterfactual.

Counterfactuals

[Slide 4] A counterfactual sentence is a conditional sentence in the subjunctive mood; typically, though not necessarily, it has the form "If it were the case that A, then it would also be the case that B", where A and B are propositions.

Symbolically, it's rendered as $A \mathrel{\square\!\!\rightarrow} B$.

As I hinted, there are of course variations in syntax: here's a famous example from David Lewis. **[Slide 5]**

A Simple CTC

You will notice that the clauses (1) and (2) of the unqualified dependence relation are each in the form of a counterfactual. So, the relation of dependence is a relation of *counterfactual* dependence. We can now replace the broad word "somewise" with "counterfactually" **[Slide 6]**:

> Where ***c*** and ***e*** are distinct events, ***e*** *counterfactually* depends on ***c*** if and only if (1) if it were the case that ***c*** occurs, then it would be the case that ***e*** occurs, and (2) if it were the case that ***c*** does not occur, then it would be the case that ***e*** does not occur.

To keep things simple, we are going to consider only events that actually do occur, so we are able to drop clause (1) and we shall do so in a moment.

However, I want now to bring in David Lewis again. Here's the great man. **[Slide 7: Picture of David Lewis]** Whereas we have linked difference-making with counterfactual dependence, Lewis, in his paper entitled "Causation" published in 1973, linked difference-making with the cause-and-effect relation. He did so in the following way **[Slide 8]**:

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it [*i.e. the difference-maker or cause*] been absent, its effects – some of them at least, and usually all – would have been absent as well. [Addition to original in italics.]

With that Lewis quotation in mind, our next move will be simply to *equate* counterfactual dependence with causal dependence **[Slide 9]**:

> Where ***c*** and ***e*** are distinct events, and both occur, ***e*** *causally* depends on ***c*** if and only if: if it were the case that ***c*** does not occur, then it would be the case that ***e*** does not occur.

The final move is to assume that causation is merely the inverse of causal dependency. So, if ***e*** causally depends on ***c***, then ***c*** is a cause of ***e***. This gives us – at long last - the final form of a Simple CTC **[Slide 10]**:

> Where ***c*** and ***e*** are distinct events, and both occur, ***c*** is a *cause* of ***e*** if and only if: if it were the case that ***c*** does not occur, then it would be the case that ***e*** does not occur.

A Problems for the Simple CTC – Early Pre-emption

Actually, no-one endorses this Simple CTC. One of the problems that it encounters is the problem of what's called early pre-emption. **[Slide 11]**

Here's a traveller who embarks on a solo journey across a desert. However, an enemy has surreptitiously put a highly virulent poison into the traveller's canteen of water. During the journey the traveller dies. On those facts, it is plausible to attribute the cause of death to the poisoning.

We can show this schematically [**Slide 12 – click as necessary**]

If you are thinking, surely the traveller's death might have occurred for some other reason, you are right. As it happened, another enemy had punctured the canteen so that the water leaked out during the journey before the traveller needed to drink and the traveller actually died of thirst. Thus, the potential death by poisoning was pre-empted by the puncturing of the canteen.

Again, we can show this schematically [**Same Slide – click further as necessary**].

So, now that we know the actual cause, we can see whether the Simple CTC actually identifies it. According to the CTC:

> The puncturing of the traveller's water canteen was a cause of his death, for if the water canteen had not been punctured, the traveller's death would not have occurred.

Unfortunately, that is false. It's false because in the absence of the puncture, the poisoning would have caused the traveller's death. One way to see it, is that the puncturing of the water canteen was sufficient to cause death, but it was not necessary. Note, too, that we have here a case of causation without causal dependence. So, the Simple CTC is basically a failure.

Lewis's 1973 CTC

Truth conditions of counterfactuals

So far, I've tacitly ignored one difficulty, which is that counterfactuals seem to be as much in need of explanation as does causation. In particular, if one is to analyse causation in terms of counterfactuals, it is necessary to give a noncircular account of the truth conditions of counterfactuals. This proved to be very challenging, despite some brave attempts by Roderick Chisholm and Nelson Goodman in the 1940s.

However, Robert Stalnaker in 1968 and David Lewis in 1973 independently proposed a method of establishing the truth conditions of a counterfactual by means of possible world semantics, which is the current orthodoxy. **[Slide 13]**

I've illustrated Lewis's approach very crudely in the slide. Say we're seeking the truth conditions at our actual world of the counterfactual "If it were the case that A, then it would be the case that C". I've labelled the actual world W. World W1 is the closest possible world where both the propositions A and C are true. World W2 is the closest possible world in which A is true but C is false. Provided that there is a possible world such as W1 that is closer to W than any possible world such as W2, as depicted in the slide, the counterfactual is deemed true. I should add that when there is no possible world in which A is true, the counterfactual is deemed to be vacuously true.

In the slide, I've shown the relation of closeness geometrically. That's' just for illustration purposes. In the theory, closeness is judged in terms of similarity to W. That's when the real complications kick in, so we'll return to Lewis and causation.

Lewis's Strategy.

In his 1973 paper entitled "Causation", Lewis formulates a relation of causal dependence along the lines of the definition we gave earlier **[Slide 14]**.

> Where ***c*** and ***e*** are distinct events, and both occur, ***e*** *causally* depends on ***c*** if and only if: if it were the case that ***c*** does not occur, then it would be the case that ***e*** does not occur.

Now, Lewis, in common with many – though not all – philosophers, believes that the relation of causation is transitive. That is to say, if A is a cause of B and B is a cause of C, then A is a cause of C. However, counterfactual dependence and therefore causal dependence, is not transitive.

Causal chains

Lewis's strategy is first to introduce the notion of a *causal chain*, which can be defined thus **[Slide 15]**:

> A causal chain is a finite sequence of distinct actual events in which each event is causally dependent upon the preceding event.

Lewis now defines causation in terms of such a causal chain, thus **[Same slide, click to fade in this sentence.]**:

> Where ***c*** and ***e*** are distinct events, ***c*** causes ***e*** if and only if ***c*** and ***e*** both occur, and there exists a causal chain leading from ***c*** to ***e***.

Technically, Lewis has defined causation not in terms of causal dependency but in terms of the *ancestral* of causal dependency, which renders causation transitive.

An advantage of the Lewis 1973 CTC

Lewis's manoeuvre also has the happy consequence that his theory overcomes the problem of early pre-emption.

This can be illustrated by our tale of the Desert Traveller. **[Slide 16]**: the lower chain is effective whereas the upper chain is *severed* by the inhibitory effect of the emptying of the water canteen. So, there is no causal chain between the poisoning and the death but there is a causal chain from the puncturing of the canteen to the death. Accordingly, unlike our Simple CTC, Lewis's 1973 theory correctly identifies the cause of the traveller's death.

Problems for the Lewis 1973 CTC

Subsequent work showed Lewis's 1973 theory to be vulnerable to several objections.

The problem of late pre-emption

[Slide 17] Billy and Suzy are a couple of street urchins much given to catapulting rocks at things. Suzy aims a rock at a glass bottle and hits it, whereupon it shatters. Billy, whose aim is as good as Suzy's also aims a rock at the bottle but his rock takes a fraction of a second longer to reach the bottle's position.

We can trace a stepwise series of events initiated by Suzy's release of her catapult and terminating in the shattering of the bottle, that is her rock being at various positions at various times during its flight through the air. However, the shattering of

the bottle does not causally depend on any of those events because if Suzy's rock had failed to reach the bottle, the shattering would still have occurred owing to the impact of Billy's rock.

The problem of symmetrical overdetermination

[Slide 18] Suzy and Billy (them again!) catapult their rocks at one and the same bottle and strike it simultaneously with sufficient force that the impact of either rock on its own would have been enough to shatter the bottle. We can make the same observations about Suzy's effort as we did in the preceding example.

{As in the preceding case, we can trace a stepwise series of events initiated by Suzy's release of her catapult and terminating in the shattering of the bottle, that is her rock being at various positions at various times during its trajectory. However, the shattering of the bottle does not causally depend on any of those events because if Suzy's rock had failed to reach the bottle, the shattering would still have occurred owing to the impact of Billy's rock.}

However, the problem is that analogous observations now apply to Billy's effort. The upshot on this analysis is that neither catapulting event was a cause of the bottle's shattering. Which seems odd.

Lewis's response (2000: 80) is that our intuitions in cases of symmetrical overdetermination are so unclear that such cases are not useful as test cases.

The problem of trumping pre-emption

[Slide19] First, here's the prince. Now the laws of magic are such that only the first spell cast on any given day is effective and that it takes effect at midnight. Spells work directly, i.e. without intermediaries. Causation at a distance, if you like. **[Click]**

One day, at 12 noon, the wizard Merlin casts a spell, the first on that day, to turn the prince into a frog. **[Click]** Later that afternoon, the wizard Morgana casts a spell, the only other spell that day, to turn the prince into a frog. At midnight the prince duly turns into a frog **[Click]**

Under the specified conditions, Merlin's spell is clearly the cause of the prince's enfrogment, whereas Morgana's is clearly not a cause. Morgana's spell is trumped by Merlin's. However, there is no counterfactual dependence of the enfrogment on Merlin's casting of the spell, since if for some arcane reason Merlin's spell didn't work, Morgana's spell would have caused the effect. So, we have a case of causation without counterfactual dependence. Furthermore, there is no chain of causal dependencies in either spell.

Lewis's 2000 CTC

In 2000, Lewis published a paper in which he presented a new CTC.

{*The paper was in fact based on a couple of lectures that Lewis presented in 1999 at Harvard University. A revised version appears in the anthology edited by John Collins, Ned Hall and Laurie Paul.}*

To explain the new theory, we need the notion of an *alteration*, as conceived by Lewis **[Slide 20]**:

> An alteration is an event, which may be actualised or unactualized, that occurs at a slightly different time or in a slightly different manner from a given event.

An alteration is – by definition - a modally very fragile event: that is to say, *any* difference in the time or manner of occurrence will give rise to a different event.

We now introduce the notion of the *influence* of one event on another. **[Slide 21]** And here's the full-dress version:

> Where ***c*** and ***e*** are distinct actual events, ***c*** *influences* ***e*** if and only if there is a substantial range c_1, c_2, … of different not-too-distant alterations of ***c*** (including the actual alteration of ***c***) and there is a range e_1, e_2 … of alterations of ***e***, at least some of which differ, such that if c_1 had occurred, e_1 would have occurred, and if c_2 had occurred, e_2 would have occurred, and so on.

Lewis's 1973 CTC considers only counterfactuals stating that *whether* one event occurs depends on *whether* another event has occurred. However, the new theory considers counterfactuals stating that whether, when and how one event occurs depends on whether, when and how another event occurs.

Causation, according to Lewis, is the ancestral of the influence relation **[Slide 22]**:

> Where ***c*** and ***e*** are distinct actual events, ***c*** causes ***e*** if and only if there is a chain of stepwise influence from ***c*** to ***e***.

An advantage of Lewis's 2000 CTC

This newer theory handles the usual cases of late pre-emption much better than the older theory.

[Slide 23] Recall the scenario wherein Suzy's rock shatters the glass bottle just before Billy's rock arrives. Consider a slight alteration in Suzy's effort, say her rock is launched half a second earlier. Then the shattering of the bottle will be altered too: it occurs half a second earlier and hence constitutes a different event. The bottle's shattering is a modally fragile event. However, if we hold Suzy's effort fixed

and make a corresponding alteration to Billy's effort, the shattering remains unchanged. (Any alterations in Billy's effort that *would* change the shattering would *not* be in accordance with the definition of 'influence' which stipulates alterations that are "not-too-distant".) Accordingly, the analysis yields the intuitively correct result that it is Suzy's effort that is the cause.

The newer theory may also cope with trumping cases. Let us return to our wizards. First consider an alteration of Merlin's spell such that it is prince-to-porcupine instead of prince-to-frog: the transmogrification at midnight would be correspondingly altered. Now hold Merlin's original spell and other conditions fixed except for an alteration of Morgana's spell to, say, prince-to-peacock. The effect of Merlin's spell would be unaltered: the prince turns into a frog as before. Accordingly, the analysis yields the intuitively correct result that it is Merlin's spell that is the cause.

A Possible Counterexample to Lewis's 2000 Theory

Lewis's 2000 theory has attracted a number of objections, but I want to present just one.

[Slide 24] Consider a main railway track that leads to a town called Clarksville. Before any train reaches Clarksville, it goes over a set of points where a branch line splits off from the main track. The points are operated by a lever. The last train to Clarksville is due to pass over the points at 6 pm. However, at 5 pm a bad guy operates the lever so that the train is directed onto the branch line. **[Same slide – Click]** I call him a bad guy because the branch line leads to a dead end, and so the train eventually derails at the end of the track. **[Same slide – click]**

It is clear that the operation of the lever is a cause of the train's derailment. However, it seems that there is no alteration in the way the lever is operated by the

bad guy that affects the way the train derails. The lever could be operated at a different time, say 4.50 or 5.10 pm or at different speeds, e.g. slowly and deliberately or with a violent shove; or by the use of a kick with a foot rather than a push with a hand. Lewis's appeal to a chain of stepwise influence doesn't help. Once the points have been reset to the branch line, there is no subsequent moment that is influenced by the resetting of the points *and* that influences the derailment.

So, this is a putative case of causation without a chain of stepwise influence.

David Hume

[Slide 25] Now, of course, no talk on causation should omit a mention of the great Scottish philosopher David Hume (1711 – 1776). Hume in his *Enquiry concerning Human Understanding* (1748) famously offered two definitions of a cause. I have given Hume's second definition. I shall end with a counterfactual: If one were to substitute the word "event" in place of Hume's word "object", then his definition would strongly resemble our Simple CTC.

That's it. Thank you. If you have any benign comments or easy questions, we have a few minutes left.

……