PART IA, INDUCTION & CAUSATION | Michaelmas 2018 | Dr M. Steenhagen
http://msteenhagen.github.io/teaching/2018inc/

Lecture 4: **The counterfactual theory of causation**

Causes as difference makers

Instead of thinking about causes as (conditionally) necessary or sufficient conditions, we can think of causes as difference makers. Also this idea seems to have been promoted by David Hume:

> The only immediate utility of all sciences, is to teach us, how to control and regulate future events by their causes. Our thoughts and enquiries are, therefore, every moment, employed about this relation: Yet so imperfect are the ideas which we form concerning it, that it is impossible to give any just definition of cause, except what is drawn from something extraneous and foreign to it. Similar objects are always conjoined with similar. Of this we have experience. Suitably to this experience, therefore, we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never existed. (*Enquiry*, VII, 2)

David Lewis develops Hume's 'second definition': 'where the first object had not been, the second never had existed'. Lewis's theory is a three-stage account. First, it exploits a more general theory of counterfactual dependence. This theory allows us to understand the relation of causal dependence. And this understanding of causal dependence in turn allows us to define causation.

Counterfactual dependence

Lewis thinks that the explanatory power of counterfactual conditionals has been underestimated. Recall, a conditional is any 'if… then…' statement or sentence. The sentence coming after the 'if' is called the antecedent, and the sentence coming after the 'then' is called the consequent. A counterfactual conditional is a conditional of the form 'if it were the case that A, then it would be the case that C'.

What makes the material conditional ($A \rightarrow C$) true is familiar from TFL. But what are the truth conditions for counterfactuals? The counterfactual conditional ($A \mathrel{\square\!\!\rightarrow} C$) is not truth-functional, according to Lewis. He uses the notion of comparative overall similarity of possible worlds to define its truth conditions:

> To begin, I take as primitive a relation of *comparative over-all* similarity among possible worlds. We may say that one world is *closer to actuality* than another if the first resembles our actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off one against another. (Lewis 1973, 559)

We easily make judgments of overall comparative similarity of items. We can create a weak ordering of items that are more and more similar than some target. It is not always obvious how we weigh the relevance of specific features to the overall similarity, but we often can rely on intuition. One world is more or less similar to another to the extent that it resembles that world in matters of particular fact and law. In what follows, an 'A-world' is a world in which A is true; likewise with C.

'$A \mathrel{\square\!\!\rightarrow} C$' is true iff either (i) there are no possible A-worlds, (ii) some A-world which is also a C-world is closer to the actual world than any A-world which is not also a C-world.

Here (i) is a situation in which the counterfactual is said to be 'vacuously true'. The important case is (ii). When the counterfactual is true, it takes less of a departure from actuality to make A and C true together than it does to make A true without C. Try applying this idea to an example (e.g. Lewis's 'if it were the case that kangaroos had no tails, then it would be the case that they would topple over').

When 'A □→ C' is true we say that *C counterfactually depends on A*. (Note, this is a relation between propositions!)

Causal dependence

Using the notion of counterfactual dependence, a relation between propositions, Lewis defines causal dependence, a relation between events. Let 'c' and 'e' be terms for events (e.g. 'the assassination', 'the first world war'). Let 'O' be a predicate of events, meaning 'occurs'. Let '¬' be negation. We can now define causal dependence:

> e causally depends on c iff (1) Oc □→ Oe and (2) ¬Oc □→ ¬Oe

If c and e are actual events, then (1) is automatically true because of the stipulation that the actual world is always the closest world to itself. Since c and e actually exist, then there is a c- and-e world which is closer to actuality than any c-and-not-e world, simply because the actual world is a c-and-e world. And in any case of causation, the cause and the effect must actually exist. Hence, clause (2) is normally taken to be the heart of the counterfactual analysis: it says that if e had not occurred, e would not have occurred. This is what it is for e to causally depend on c.

Note, there are many cases of counterfactual dependence which are not cases of causation (see Kim's paper in the Sosa & Tooley volume for some examples). Lewis himself brings out that the laws of motion in a world may counterfactually depend on the laws of gravity in that world, but the latter don't cause the former (better to say: laws of motion *supervene* on laws of gravity).

But even if we limit ourselves to events we should be careful. If you park your car on a double yellow line, then you break the law. But parking the car there doesn't cause you to break the law (it constitutes breaking the law in this instance). What we must add to exclude cases like this is to say that the events related as cause and effect must be *numerically distinct* from one another.

Causation

Causal dependence is a sufficient condition for causation. But it is not necessary. This is because causation is a *transitive* relation and causal dependence is not.

> Causal dependence among actual events implies causation. If c and e are two actual events such that e would not have occurred without c, then c is the cause of e. But I reject the converse. Causation must always be transitive; causal dependence may not be; so there can be causation without causal dependence. Let c, d, and e be three actual events such that d would not have occurred without c and e would not have occurred without d. Then c is a cause of e even if e would still have occurred (otherwise caused) without c. (Lewis 1973, 563)

A relation R is transitive when 'aRb' and 'bRc' imply 'aRc'. A relation is non-transitive when this is not the case; a relation is intransitive when 'aRb' and 'bRc' imply 'not-aRb'.

Causation is defined in terms of a *chain* of counterfactual dependence. A causal chain is defined as a sequence of actual events, c, d, e... etc., where d depends on c, e depends on d etc. Then c is a cause of e when there is a causal chain from c to e. Example:

> Suppose I shoot the president, and this brings about a revolution, which in turn brings about the president's rival ascending to power. Let's suppose that each later stage in this process causally depends on the earlier stage. Lewis would say that even though it is true that my act caused the president's rival to ascend to power, it need not thereby be true that the president's rival's ascent is causally dependent on my shooting, since it need not be true that in the closest world in which I did not shoot, he did not ascend to power (maybe the whole situation is so politically unstable that someone else would have shot if I hadn't). So we have causation between my action and the eventual outcome without causal dependence between my action and the eventual outcome.

A relation R^* that is constituted by a chain of relations R is called the *ancestral* of R (cf. Frege, *Begriffschrift*). The ancestral of a relation R is that relation which stands to R as the relation of being an ancestor stands to the relation of being a parent. The relation 'ancestor' can be roughly defined as follows: x is an ancestor of y iff x is a parent of y, or x is a parent of a parent of y, or x is a

parent of a parent of a parent of *y*... and so on. (As we could say '*x* is an ancestor of *y* in the Parent-series'.) While 'x is a parent of y' is not transitive, 'x is an ancestor of y' is.

The same structure holds for the relations 'x causally depends on y' and 'x is a cause of y'. Causation is the ancestral of the relation of causal dependence.

The problem of redundant causation

Something seems a case of redundant causation when it is obvious that c causes e, but it is not true that if c had not occurred, e would not have occurred (i.e. there seems to be causation without causal dependence). Two different kinds of redundant causation are central:

1. **Overdetermination**: e *actually* has two independent causes, c1 and c2. Since c1 and c2 are independent, either would occur without the other. In the closest world where c1 does not occur, e still occurs, because c2 brings it about. And vice-versa with c2. So e causally depends on neither c1 nor c2.
   Example: two independent assassins shoot and kill the tyrant. Arguably in such a case, if the first assassin had not fired, the second would still have killed the tyrant. And the same applies to the second.
2. **Pre-emption**: c actually causes e, but there is a 'back-up' cause waiting to cause e if c fails. So e does not causally depend on c.
   Example: one assassin shoots and kills the tyrant. But another is waiting to shoot just in case the first one misses. The first one doesn't miss, so his shot is the cause of the death. But it isn't true that if the first had missed, the tyrant would not have died.

Lewis can respond to overdetermination by pointing out that we have no good reason to treat these events as distinct. If these occurrences are exactly simultaneous, then they seem to be (parts of) the same event.

Lewis can deal with standard pre-emption by appealing to his definition of causation in terms of a chain of causally dependent events. The first assassin's shot causes the tyrant's death because there is a chain of events between this shooting and the death. Each event in this chain is causally dependent on the one before, but this does not mean that the death is causally dependent on the shooting. So the fact that there is a back-up does not stop the shot causing the death.

Late pre-emption

However, we should distinguish early pre-emption and late pre-emption:

> **Early pre-emption:** c causes e, but there is a 'back-up' cause waiting to cause e if c fails to happen. So e does not causally depend on c. Here the pre-empting cause is imagined to occur well before e, so that there is a causal chain of events between the pre-empted cause and e.
>
> **Late pre-emption:** c causes e, but there is a second, parallel process that would have caused e if c or any of its intermediary effects c*, c'... etc. had failed. Here the pre-empting cause is imagined to occur immediately before c, so that there is no causal chain of events between the pre-empted cause and e.

In *Causation: A Users Guide,* L.A. Paul and Ned Hall discuss examples of late preemption. Here's a clear case:

> Suzy and Billy, two friends, both throw rocks at a bottle. Suzy is quicker, and consequently it is her rock, and not Billy's, that breaks the bottle. But Billy, though not as fast, is just as accurate: Had Suzy not thrown, or had her rock somehow been interrupted mid-flight, Billy's rock would have broken the bottle moments later. This case, like [cases of early pre-emption], features a cause of an event that is accompanied by a backup, sufficient to bring about the effect in the cause's absence. But unlike those cases, the actual cause fails either to interrupt the backup process, or to interact with it in such a way as to render it insufficient for the effect.

The counterfactual theory of causation is equipped to deal with early pre-emption, but has difficulties with late pre-emption. (Lewis responds to late pre-emption cases by highlighting that events are 'modally fragile', so that if the pre-empting cause had its way, a slightly different event would have been the result, which renders it false that the actual effect would have occurred had the actual cause not happened. How satisfactory is this reply?)