Lecture 10: Causation 1

1. Introduction
2. Recap on Counterfactuals
3. The Simple Counterfactual Account of Causation
4. Backtracking
5. Redundant Causation
6. Transitivity

Lecture 10: Causation 1

1. Introduction

- Examples of causal statements:
  - (1) 'Alexander Litvinenko's death was caused by the ingestion of a large dose of Polonium 210'
  - (2) 'You always get lost because you just can't map-read.'
  - (3) 'The Iraq war caused a surge of sympathy for terrorist causes.'
  - (4) 'Hamdi broke the window by kicking his toy lorry into it.'
- Causal notions are also widely used in philosophical analyses of *all sorts* of concepts. We get causal analyses of:
  - (i) knowledge,
  - (ii) mental content / word meaning,
  - (iii) moral / legal responsibility
  - (iv) rational decisions
  - (v) etc…

Lecture 10: Causation 1

1. Introduction

- Despite our familiarity with causal concepts, it turns out that, upon reflection, we find it *extremely* difficult to say pretty much anything uncontroversial about what exactly causation *involves*. (in this respect, the situation is very much similar to modal notions)
- Philosophers have poured a *huge* amount of effort into the analysis of causal statements, with so far relatively little agreement as to how what the proper analysis should look like.
- This, however, isn't to say that we haven't come a long way since early work on the subject.
- In what follows, I won't be commenting on this philosophical journey. I will depart from the standard historical presentation of work on causation (starting with Hume, through Russell, Mackie and so on) and cut straight to one of the *currently* most popular analyses: the counterfactual analysis.

Lecture 10: Causation 1

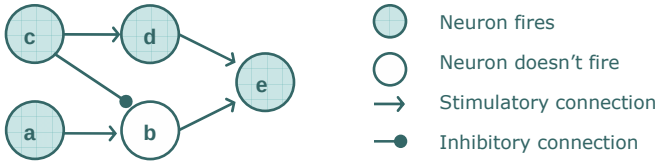
1. Introduction

- There are a number of reasons for this choice:
  - (i) the counterfactual analysis is probably the single most widely discussed approach to the analysis of causation,
  - (ii) it is possibly the most successful,
  - (iii) it enables us to put our lecture on counterfactuals to good use
- Warning: the literature on causation in general, and counterfactual analyses of causation in particular, is a never-ending succession of analyses followed by counterexamples.
- The analyses and counterexamples start off fairly simple but very quickly head towards the increasingly baroque and incomprehensible.
- Some find this fun, some find it tedious.
- Whatever the case, this is a pattern *typical* of much analytic philosophy, so this should give you an interesting insight into the way the discipline is practiced.

Lecture 10: Causation 1

1. Introduction

▪ To help with the exposition of the counterexamples, I will make use of so-called 'neuron diagrams', which provide an easy way to visualise various causal situations:



▪ Temporal order is represented by spatial order: leftmost events occur first.

▪ What the diagram describes here, is a situation in which (i) c and a both fire, (ii) c's firing causes d to fire, which in turn causes e to fire, and (iii) a's firing however fails to bring about e's firing due to the fact that c's firing prevents it from bringing about b's firing, which would have caused e's firing.

Lecture 10: Causation 1

2. Recap on Counterfactuals

▪ Before we get started, a quick recap on counterfactuals…

▪ Counterfactual conditionals are natural language expressions of the form 'had it been the case that p, it would have been the case that q' (formally: $p \mathrel{\square\!\!\rightarrow} q$).

▪ We have seen that it is now widely agreed that counterfactual conditionals correspond neither to the logician's material conditional ($p \rightarrow q$, or again $p \supset q$) or to the so-called 'strict' conditional ($\square(p \rightarrow q)$).

▪ The main motivation for this was that it would appear that whilst the aforementioned conditionals satisfy the following three principles, counterfactual conditionals appear to satisfy none of them (here, pRq stands in for a dummy relation that can correspond to either a material, a strict or a counterfactual conditional):

(i) Contraposability: if pRq then $(\sim q)R(\sim p)$.

(ii) Monotonicity: if pRr then $(p \& q)Rr$.

(iii) Transitivity: if pRq and qRr then pRr.

Lecture 10: Causation 1

2. Recap on Counterfactuals

- We the examined two attempts to account for the semantics of counterfactuals in terms of possible worlds: Lewis' and Stalnaker's.
- I suggested that, Lewis' account was preferable to Stalnaker's for various fairly technical reasons (falsity of Conditional Excluded Middle, etc.).
- Here is the account in question, which we called [L], for 'Lewis':

  [L] p □→ q iff there is at least one p&q-world closer to the actual world than any p&~q-world.
- Ok then, back to business...

Lecture 10: Causation 1

3. The Simple Counterfactual Account of Causation

- Let's start simple:

  CTC_1: c caused e iff (i) c, (ii) e, and (iii) ~c □→ ~e
- So, for example, my turning the key in the ignition caused my car to start this morning because: (i) I turned the key in the ignition, (ii) my car started, and (iii) had I not turned the key in the ignition, my car wouldn't have started.
- (iii), on Lewis' account of counterfactuals, comes out as: there is a possible world in which I don't turn the key in the ignition and the car doesn't start that is closer to actuality than any possible world in which I don't turn the key and the car *does* start.

Lecture 10: Causation 1

3. The Simple Counterfactual Account of Causation

- CTC_1 faces an immediate problem, according Kim (1973). The following counterfactuals are true:

  (5) 'Had I not come to the lecture, I would have not have come to the lecture.'

  (6) 'Had I not written "counterfac", I wouldn't have written "counterfactuals"'

- It would seem that, according to CTC_1 we have to say that: (i) my coming to the lecture caused my coming to the lecture, and (ii) my writing "counterfac" caused me to write "counterfactuals".
- But this seems wrong: surely in neither case do we have causation (my coming to the lecture just *is* my coming to the lecture and my writing "counterfac" is just *part* of my writing "counterfactuals": things don't cause themselves and parts don't cause wholes).
- Thankfully, CTC_1 is easy to patch up here:

  CTC_2: c caused e iff (i) c, (ii) e, (iii) $\sim c \; \Box\!\!\rightarrow \sim e$ and (iv) c and e are entirely distinct.

Lecture 10: Causation 1

4. Backtracking

- But there is a potential problem here (known as the 'problem of effects')…
- Say that I have a button rigged up to a bomb. There is no other way of blowing up the bomb without pressing the button. I press the button. The bomb goes off. The following counterfactual is true:

  (7) 'Had I not pressed the button, the bomb wouldn't have exploded.'

- CTC_2 correctly tells me that my pressing the button caused the bomb to go off.
- But consider… Can't we also asset another counterfactual?

  (8) 'Had the bomb not exploded, it would have had to have been the case that I hadn't pressed the button.'

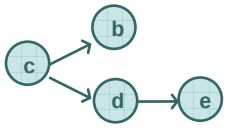
- So shouldn't we *also* say that the bomb's explosion caused the button to be pressed?
- One easy fix would be to move to:

  CTC_3: c caused e iff (i) c, (ii) e, (iii) $\sim c \; \Box\!\!\rightarrow \sim e$, (iv) c and e are entirely distinct and (v) c occurred before e.

Lecture 10: Causation 1

4. Backtracking

▪ This works for the 'problem of effects', but issues remain. Consider the same scenario but add the fact that a small warning light (b) goes on immediately after the button (c) is pressed but before the bomb is detonated (e).



▪ Couldn't we say here:

(9) 'Had the light not gone on, it would have had to have been the case that the bomb wasn't going to explode.' ?

▪ And if we can say *that*, don't we end up with the result that the light's going on caused the bomb to detonate, despite the fact that the former occurred earlier than the latter? (This problem is known as the 'problem of epiphenomena'.)

Lecture 10: Causation 1

4. Backtracking

▪ The key to addressing both worries (i.e. the problem of effects and the problem of epiphenomena) is to pay attention to the grammatical form of the problematic counterfactuals.

▪ Whilst it seems ok to assert:

(9) 'Had the light not gone on, it would have had to have been the case that the bomb wasn't going to explode.'

but it seems false to say:

(10) 'Had the light not gone on, the bomb wouldn't have exploded.'

▪ In other words, there are in fact *two* types of counterfactuals with different truth-conditions: 'backtracking' counterfactuals, such as (9) and 'non-backtracking' counterfactuals such as (10) (this is Lewis's terminology, but it has now become standard).

▪ Let $\Box\!\!\rightarrow_b$ stand for the backtracking counterfactual, and $\Box\!\!\rightarrow_{\not b}$ stand for its non-backtracking counterpart

Lecture 10: Causation 1

4. Backtracking

- The counterfactualist argues that if we understand condition (iii) of:

  CTC_2: c caused e iff (i) c, (ii) e, (iii) $\sim c \; \Box\!\!\rightarrow \sim e$ and (iv) c and e are entirely distinct.

as $\sim c \; \Box\!\!\rightarrow_{nb} \sim e$, we get the right results with regards to our two problems.

- How does Lewis's theory of counterfactuals account for the fact that (9) is true and (10) is false? Well he needs to claim that there are differences in interpreting the notion of 'closeness' from one form to the other.

Lecture 10: Causation 1

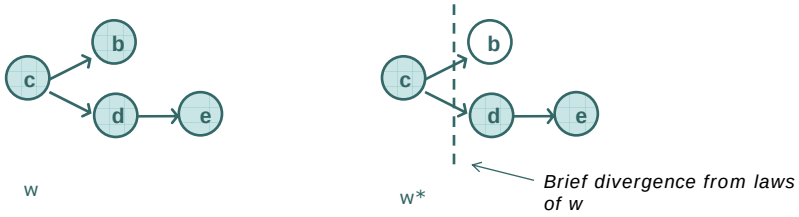
4. Backtracking

- For non-backtracking counterfactuals of the form $\sim p \; \Box\!\!\rightarrow_{nb} \sim q$, it is standard to stipulate that the 'closest' possible $\sim p$-world w* to the actual world w is such that:
  - w* is indistinguishable from w, both in terms of laws and particular facts, up to shortly before the time t that p occurs in w,
  - at this point, a very minor departure from the laws of w, in w*, prevents p from occurring.
  - from t onwards, the laws of w reapply in w* and business proceeds as usual.
- For backtracking counterfactuals, things are less clearly understood. One possibility would be to stipulate that the relevant 'closest' possible $\sim p$-world w** to the actual world w is such that w** is identical to w in terms of laws and as close as possible to w, given the absence of p.

Lecture 10: Causation 1

4. Backtracking

▪ Now see how this yields the right result for our bomb and light example, if we take the non-backtracking reading of the relevant counterfactual…

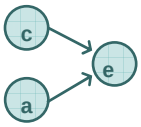


▪ The 'closest' possible world in which the light doesn't come on is a world in which everything is identical to the actual world just prior to the time of the light going on, by which time the process leading to the explosion of the bomb is already on its way and the bomb goes off.

Lecture 10: Causation 1

5. Redundant Causation

▪ Problem (1): Overdetermination.



▪ Neurons c and a fire. The signals reach e simultaneously. e fires.

▪ According to CTC_2, neither c's firing, nor a's firing caused e's firing, because: (i) $\sim c \; \Box\!\!\rightarrow_{hb} e$ and $\sim a \; \Box\!\!\rightarrow_{hb} e$.

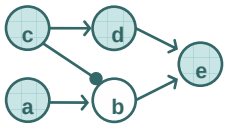
▪ This seems to many like the wrong result: shouldn't we say that *both* c and a's firings caused e's firing?

▪ Lewis (1973: 567), however, complains here that we do not have clear intuitions as to the causal structure of the example (do you agree?); but let's grant him this for sake of argument because there is more trouble to come…

Lecture 10: Causation 1

5. Redundant Causation

- Problem (2): Early Preemption.



- This is the diagram that I showed you in the introduction.
- c fires, causing d to fire, which in turn causes e to fire. a's firing never gets to bring about anything, as c's firing prevents it from bringing about B's firing.
- Again, according to CTC_2, neither c's firing nor a's firing caused e's firing, because: (i) $\sim c \; \Box\!\!\rightarrow_{hb} e$ and $\sim a \; \Box\!\!\rightarrow_{hb} e$.
- This again seems like the wrong result: surely it was c's firing that brought about e's firing.

Lecture 10: Causation 1

5. Redundant Causation

- Lewis offers a variant on CTC_2 that handles the early preemption case (but not the overdetermination case).
- He suggests that we identify causation not with counterfactual dependence, but with the *ancestral* of counterfactual dependence:

  CTC_4: c caused e iff (i) c, (ii) e, (iii) c and e are entirely distinct, and (iv) there is an actual finite chain of counterfactually dependent events leading from c to e (i.e. a set of actual events $S = \{d_1, \ldots, d_n\}$, such that $\sim c \; \Box\!\!\rightarrow_{hb} \sim d_1$, $\sim d_1 \; \Box\!\!\rightarrow_{hb} \sim d_2$, …, $\sim d_{n-1} \; \Box\!\!\rightarrow_{hb} \sim d_n$, $\sim d_n \; \Box\!\!\rightarrow_{hb} \sim e$).

(of course, the proposed amendment can provide a solution to our problem only if counterfactual dependence is non-transitive: if counterfactual dependence is transitive, condition (iv) entails $\sim c \; \Box\!\!\rightarrow_{hb} \sim e$, and we are back in trouble with early preemption)

Lecture 10: Causation 1

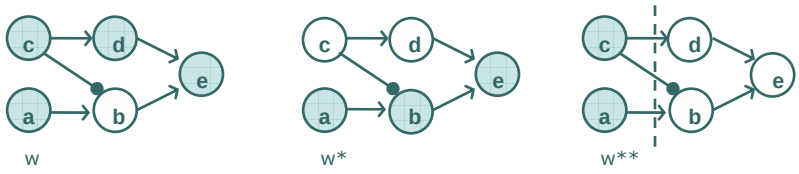
5. Redundant Causation

- See how this applies to the early preemption case:

  (i) had it not been the case that c fired, it wouldn't have been the case that d fired. (see w*)

  (ii) had it not been the case that d fired (assuming that the relevant counterfactuals don't backtrack), it wouldn't have been the case that e fired (because we hold the world fixed until shortly before d fires, and by then c has already sent its inhibitory signal towards b, so b still wouldn't have fired). (see w**)

- So the conditions are met for c's firing to cause e's firing. Not so for a's firing.



Lecture 10: Causation 1

5. Redundant Causation

- Intuitively, the way that Lewis's amendment works is that in the early preemption case, there is a point in time at which the signal from the preempted non-cause (i.e. a's firing) is cut short and the signal from the actual cause (i.e. c's firing) finds itself with no backup.
- We will see next week that we can construct cases of preemption that don't offer this way out to Lewis.

Lecture 10: Causation 1

6. Transitivity

- An interesting thing about CTC_4 is that it does justice to our intuitions that causation is transitive, i.e. that if c caused d and d caused e, then c caused e.
- For instance, if the interview panel were unfriendly towards Mr. Griffiths and this caused him to lose his nerve, which in turn cost him the job, it presumably follows that the unfriendliness of the panel was a causal factor in Griffiths' not getting the job.
- According to CTC_4, causation is transitive. To see why,
  - Consider a chain of actual events c, then i_1 then d, such that $\sim c \; \Box\!\!\rightarrow_{hb} \sim i_1$ and $\sim i_1 \; \Box\!\!\rightarrow_{hb} \sim d$. According to CTC_4 we have c caused d.
  - Now consider another chain, involving d, then i_2, then e, such that $\sim d \; \Box\!\!\rightarrow \sim i_2$ and $\sim i_2 \; \Box\!\!\rightarrow_{hb} \sim e$. According to CTC_4 we have d caused e.
  - But of course, we now also have a chain of actual counterfactually dependent events leading from c to e (c, i_1, d, i_2, e) and hence we also have, according to CTC_4, c caused e.

Lecture 10: Causation 1

6. Transitivity

- So according to CTC_4, causation is transitive.
- Note that, *assuming* that the relation of counterfactual dependence *isn't* transitive (which some – albeit not many - deny), this *wasn't* the case for the earlier versions that I presented.
- Why? Because, *obviously*, if counterfactual dependence isn't transitive and causation amounts to counterfactual dependence, causation comes out non-transitive as well.
- I have presented transitivity of causation as a good result for CTC_4. However, some people have argued that it is in fact a *bad* result (see Hall (2000) or Lewis (2000)).

Lecture 10: Causation 1

6. Transitivity

- Consider the following situation (from Lewis 2000): White and Black are playing a game of chess. Black makes a move, which, if not countered, would have put Black in an extremely strong position. Seeing this, White counters, and as a result of this counter, heads to victory.
- Here, some claim: (i) Black's move caused White's move, (ii) White's move caused White's victory, but (iii) Black's move didn't cause White's victory. If this is so, transitivity fails for causation.
- Another example (perhaps more convincing): Agent Smith pulls the trigger at Neo at time t-1. Neo ducks at time t, avoiding the bullet. Neo is alive at t+1.
- Here, some would claim: (i) Agent Smith's pulling the trigger at t-1 caused Neo to duck at t, (ii) Neo's ducking at t caused Neo's being alive at t+1, but (iii) Agent Smith's pulling the trigger at t-1 didn't cause Neo's being alive at t+1. Again, transitivity seems to fail for causation.

Lecture 10: Causation 1

6. Transitivity

- What does Lewis say here? See Lewis (2000: 194-195).
- I will try to say something more about transitivity next time.
- But note that whatever the fact of the matter about transitivity, there remain some awkward problems for Lewis's CTC_4... More on this next week.

Lecture 10: Causation 1

Reference

- Hall, N. (2000), 'Causation and the Price of Transitivity', *Journal of Philosophy* 97: 198-222.
- Kim, J. (1973) 'Causes and Counterfactuals', *Journal of Philosophy* 70:570-572.
- Lewis, D. (1973) 'Causation', *Journal of Philosophy* 70: 556-67.
- Lewis, D. (2000) 'Causation as Influence', *Journal of Philosophy* 97: 182-97.

Lecture 10: Causation 1

Next week... Causation 2

- Set reading: finish off Lewis's Causation + Postscript to 'Causation' (on WebCT).
- If you have time: take a look at Lewis's 'Causation as Influence' (also on WebCT).