Notes from Class Discussion: Week 2 (led by Ben Escoto)

Branden Fitelson

January 17, 2002

These are my reflections on class discussion from 1/17/02 concerning the material for week 2 (counterfactuals and counterfactual theories of causation), as led (excellently) by Ben Escoto. Please send me any corrections or additions (as my memory is less than perfect).

1 Lewis' Theory of Counterfactuals

Ben quickly reviewed Lewis' semantics for counterfactuals. Roughly, a counterfactual $p \boxrightarrow q$ (if p were true then q would have been true) is true on Lewis' account just in case all of the p-worlds which are closest to the actual world are also q-worlds. So, the key to Lewis' account of counterfactuals will be judgments of overall similarity between possible worlds. And, on Lewis' theory, we have the following ranking of importance for various world-differences:

1. Most important are differences involving widespread and diverse violations of laws of nature (big miracles). Such differences are worth 4 world distance units (wdu).

2. Second are differences leading to a lack of perfect matching of particular fact over a large region os spacetime. These are worth 3 wdu.

3. Third are differences involving localized, small, or simple violations of laws of nature (small miracles). These are worth 2 wdu.

4. The least important differences are those involving a lack of approximate similarity regarding particular fact (worth 1 wdu).

As Ben nicely explained, Lewis uses this ranking to argue that – contrary to what several authors have claimed – Lewis' theory gives the correct answer concerning the counterfactual "If Nixon had pushed the button (p) then there would have been a nuclear holocaust (q)." Several authors have argued that Lewis' theory should say that $p \boxrightarrow q$ is false, since (intuitively) the closest p-worlds are *not* q-worlds. Lewis responds by arguing that this "intuition" is incorrect, if one adopts his ranking of world-differences (as explained above). Lewis asks us to compare the actual world w with (i) the closest $p\,\&\,q$-world w', and (ii) the closest $p\,\&\,\neg q$-world w'' — with respect to (1) – (4), above.

world	p	q	(1)-diffs	(2)-diffs	(3)-diffs	(4)-diffs	overall distance from w (wdu)
w	F	F	0	0	0	0	$(0 \times 4) + (0 \times 3) + (0 \times 2) + (0 \times 1) = 0$
w'	T	T	0	1	1	1	$(0 \times 4) + (1 \times 3) + (1 \times 2) + (1 \times 1) = \mathbf{6}$
w''	T	F	0	1	2	0	$(0 \times 4) + (1 \times 3) + (2 \times 2) + (0 \times 1) = \mathbf{7}$

There are two keys to Lewis's argument. First, the closest $p\&\neg q$-world w'' requires 2 small miracles (the one leading to Nixon pushing the button and the one leading to the apparatus failing, so that no detonation occurs), whereas the closest $p\,\&\,q$-world w' requires *only one* small miracle (the one leading to Nixon pushing the button). Second, notice how crucial the difference ranking is. If (3) and (4) were switched, then the judgment that w' is closer to w than w'' would no longer be correct. It is essential to Lewis' argument that a lack of approximate similarity regarding particular fact be

less important than a greater number of small miracles. If this were reversed, then we would have w' being 6 wdu away from w, and w'' being only 5 wdu away from w, which would be bad news for Lewis' theory.

Judgments of overall similarity (in general) are notoriously difficult. There are many arguments in the literature which make the case that such judgments are inherently *contextual* or *indexical*. When someone asks whether w' is more similar to w than w'' is, one is tempted to respond with a request for clarification. One is tempted to ask "similar *in which respects*"? Any two (distinct) things will share many properties and fail to share many properties. How do we weight these differences to come up with a judgment of *overall* similarity? Lewis claims to be able to provide us with a ranking that does the job for similarity comparisons between possible worlds (for the purposes of applying his theory of counterfactuals), but is this ordering/weighting objectively correct? Lewis's ordering does seem to allow him to provide "intuitive answers" concerning the truth of counterfactuals. But, it seems that he has simply "reverse engineered" his proposed ranking for this purpose. Are there *independent grounds* for thinking that this ranking is correct? If we are stuck with the contextuality of judgments of overall similarity, then it seems we are also stuck with the contextuality of counterfactual dependence (in Lewis's sense), and therefore contextuality of causation *qua* counterfactual dependence.[1] I will return to the contextuality issue, below.

Ben turned Scot's Holmes example into a counterexample to Lewis' theory of counterfactuals. Here is Ben's rendition of the example:

> Tuesday 1:00AM Holmes saves the world
> Tuesday 1:05AM Holmes dies
>
> On Lewis' theory, "If Holmes had had a heart attack on Tuesday, the world would (still) be safe" is true, because in the nearest world the heart attack occurs late (because more of the past is exactly similar, the the future is somewhat similar). Counterfactuals like "If Holmes had a heart attack on Tuesday, it would have been between 1:00AM and 1:05AM" would also still seem to be true (which seems unintuitive).

2 The Contextuality of Causal Judgments

Both Darren and Claudia (in their postings and comments last week) raised the issue of the contextuality of causal claims and judgments. Darren seemed to suggest that any event could be considered "part of a cause of" any other event, and that it's merely a pragmatic matter as to which of these "parts" we choose to focus on and call "causes". Claudia emphasized the contextuality of causal judgment. She noted that in some contexts, we are happy to call x a cause of y, but in other contexts, we may not want to make such a claim. She raised examples involving legal/moral responsibility which emphasize the importance of context in certain kinds of "apportioning of responsibility" judgments. In class, Patrick had an interesting reply to this. He suggested that these intuitions may conflate explanatory judgments with causal judgments. Philosophers such as Van Fraassen have emphasized the contextuality (and pragmatic nature) of *explanation*. Which things we take to be explanatory do often seem to depend on contextual factors and (pragmatic) judgments of *relevance*. But, do we want to say the same thing about causal judgments? It's not so obvious to me that there will be ab asymmetry between causal and explanatory judgments. Many authors have taken explanation to be inherently bound up with causation. In any event, this is an important distinction to make. I urge you to read the pair of papers on explanation and causation by Woodward and Hitchcock. These two papers are now posted on the `syllabus page` (under optional readings for week 3). You can download these papers `here` and `here`.

[1] See Bowie's paper ''The Similarity Approach to Counterfactuals: Some Problems" for a discussion of Lewis' similarity approach to counterfactuals, and some of its less desirable features, stemming from "overall similairty".

In class, we contrasted (more formally) contextual and non-contextual accounts of causation. Roughly, non-contextual accounts of causation will take the causal relation C to be a binary relation (between events, or event types, etc.), whereas, contextual accounts will take C to be a many-placed relation. In our reconstruction of Claudia's suggestions, we arrived at a five-place contextual causal relation. According to this account, causal claims will be (roughly) of the form "a — as opposed to $x_1, x_2, \ldots, x_n \in \mathcal{X}$ — causes b, relative to causal field $\mathcal{F}$ and according to causal relevance relation $\mathcal{R}$." The x_i are alternative possible causes in the context, which form a *contrast class* $\mathcal{X}$. The "causal field" $\mathcal{F}$ is the set of factors which are to be held-fixed in the assessment of a's efficacy for b. And the relation $\mathcal{R}$ is a relevance relation, which is presumably determined by contextual factors. Simplifying things a bit, we may say that the "context" $\mathcal{C}$ consists of the contrast class $\mathcal{X}$, the causal field $\mathcal{F}$, and the relevance relation $\mathcal{R}$. And, we may then abbreviate the contextual causal claim as "a causes b in $\mathcal{C}$," where this is taken as shorthand for the longer (more accurate) claim above. Contextual accounts of causation such as these have appeared in the literature. A good place to begin would be Menzies' recent paper `''Difference-Making in Context''` which appears in the new volume *Causation and Counterfactuals* edited by Collins, Hall, and Paul.

We can think of the objections raised to Lewis' "similarity" approach to counterfactuals as exposing a kind of contextuality in Lewis' account of causation. If similarity judgments are contextual, then accounts of causation (like Lewis') which depend on similarity judgments will also be contextual. Understood in this way, are there analogues of $\mathcal{X}$, $\mathcal{F}$, and $\mathcal{R}$ in Lewis' theory?

3 Lewis' Theories of Causation

3.1 The Old Theory

On Lewis' old theory, causation is simply understood as the ancestral (the transitive closure) of the counterfactual dependence relation. That is, a causes b just in case either $\neg a \boxrightarrow \neg b$ or there is a chain of counterfactual dependence between a and b (i.e., $\neg a \boxrightarrow c_1 \boxrightarrow c_2 \boxrightarrow \cdots \boxrightarrow c_n \boxrightarrow \neg b$). This theory faced two main sorts of objections involving cases of redundant causation: preemption and overdetermination. I refer the reader to Menzies' discussion in his SEP entry `''Counterfactual Theories of Causation''` for details (which was close to the discussion we had in class).

3.2 The New Theory

On Lewis' new theory, we are asked to consider "alterations" of events a and b. For instance, let the effect be a man's death. This effect can be altered in various ways (*e.g.*, the time or manner of the death). And, (intuitively) the cause of the man's death may have been ingesting poison. And, this event can be altered in various ways. For instance, he may have ingested poison on a full stomach. According to Lewis' new theory, a will be considered a cause of b (roughly) if there is a "chain of stepwise influence" from a to b. More precisely, a caused b if there is a substantial range of a_1, a_2, ... a_n of different not-too-distant alterations of a (including the actual alteration of a) and there is a range of b_1, b_2, ... b_n of alterations of b, at least some of which differ, such that $a_1 \boxrightarrow b_1$, $a_2 \boxrightarrow b_2$, ..., $a_n \boxrightarrow b_n$.

The new theory seems to handle cases of pre-emption pretty well (see Menzies). But, the new theory seems to general new examples of spurious causation. On the new theory, many events which merely change the time, manner, etc. in which the effect occurs will be counted as causes of the event. For instance, a man's eating a large meal (before ingesting poison) will be considered a cause of his death — even though the meal only delays the death, without changing the way in which the death occurs, etc.