

Free Will III: 'Could have done otherwise'

COULD

How should we understand 'could' statements? Linguistic orthodoxy now follows Angelika Kratzer's view that 'could' sentences should be read as 'could in view of ...' i.e. *relative to a set of facts or considerations*. Compare: Could I get to London by two o'clock? No, relative to some facts (scheduled rail services; reasonable cost; organizational difficulty); yes, relative to others (assistance from the RAF; constraints of the speed of light). Apply this to the time travel case discussed by David Lewis ('Paradoxes of Time Travel'): Tim has travelled back in time to a date before his father was even conceived, and is trying to kill his father's father. Could he have succeeded? We know he *doesn't* kill him since his grandfather lived on to have his father; but *could* he? 'Yes', relative to one set of considerations: he has a gun, is a good shot, no one is in the way etc.; 'No' relative to others: Tim is there, and so his grandfather must have survived in order to have fathered him. And now we say much the same about determinism more generally. A tree falls and narrowly misses a house. Was it dangerous? 'Yes', because it could easily have hit the house, even though, 'No' since, relative to the full set of facts it was determined not to. Could you have stayed in bed this morning till 12 o'clock? 'Yes', relative to one set of facts (no one was forcing you up); no, relative to another (you got here on time, or close enough). Note this isn't an *ad hoc* move for the free will debate; it is part of a general account of our modal terms.

COULD HAVE DONE OTHERWISE

How do we determine the relevant considerations for questions of whether agents could do otherwise? The standard incompatibilist claim is that *everything* up to the time of action is relevant: all the facts, laws of nature. But relative to this, no non-actual possibility statement is ever true. In general someone who insisted on talking this way would be missing important facts about the world. It matters that some things are possible and others not: in the earlier example, the tree was dangerous, even though it missed the house and did no harm. Some compatibilists want any set of considerations that will make us come out as free (List at times seems to talk this way). But there should be some constraint: what do we take the *relevant* considerations to be in our ordinary usage?

A question that helps with this: what is the experience of agency and of free will? Kushnir, Gopnik et al. 'Developing Intuitions About Free Will': at the age of four children generally think that they and others can do nothing to resist their desires (they cannot stop themselves from acting in the presence of strong ones, or bring themselves to act in their absence); by six they have generally arrived at the idea that they can. So we might say: our actions are free from our desires.

First central idea: when people believe that they can or could do otherwise, they mean that, *relative to their personal level psychological states* (most centrally desires, but also perhaps beliefs, perhaps prior intentions), they could do otherwise; that is, that it is possible to *hold those psychological states fixed* and still have them do otherwise. And this is independent of whether, *relative to the entire micro-physical structure of the universe*, they could do otherwise.

METAPHYSICS

Metaphysical problem: how can we hold that the action is possible relative to the psychological considerations, and not to the microphysical, if we accept physicalism? For doesn't fixing the physical entail fixing the psychological?

Clearly there is no way out if the psychological states are *identical* with the microphysical. If they are identical then Leibnitz's law applies: the psychological states have the same properties as the microphysical ones, including the same causal properties.

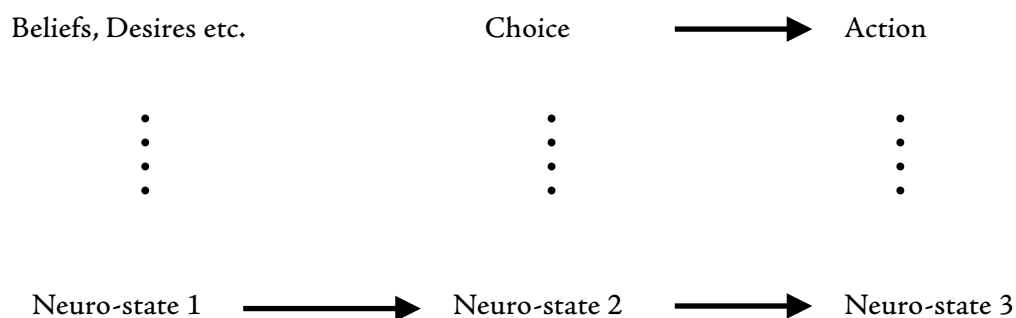
So the account, if physicalist, will have to be *non-reductive*. That is, it will have to deny that the psychological states are *identical* with the microphysical. Let us assume then a weaker claim, that the psychological *supervenes* on the microphysical, or the 'neuro-states' as we'll call them, where we understand this as saying that there could be no change in the psychological without a corresponding change in the neuro-states (in that sense, a picture on an lcd screen supervenes on the arrangement of the pixels). If B supervenes on A, we'll say that A *underlies* B (some people talk in terms of 'subvenience', but that gets clumsy).

One feature of supervenient non-reductive accounts is that they allow multiple realization: there can be many different ways of realising a given psychological state at the neuro-level (just as there may be different ways of realizing the same picture with pixels). That is a feature that we want independently: the same mental state can be realized differently in different individuals, or in the same individual at different times (e.g. after a stroke).

The we can ask whether there is a consistent account in which the following conditions obtain:

- (i) at the psychological level: prior psychological states (beliefs and desires) do not cause action, but decisions do;
- (ii) at the microphysical level: neuro-states on which prior psychological states (beliefs and desires) supervene cause neuro-states on which decision supervenes, which in turn cause neuro-states on which action supervenes; so by transitivity of causation, neuro-states on which prior psychological states (beliefs and desires) supervene cause neuro-states on which action supervenes.

Putting this as a picture (in which solid lines represent causation, dotted lines represent supervenience):



To understand how we might have something of this form, we need an account of causation. Assume that it is characterized by the conjunction of a necessity condition and a sufficiency condition. These can be understood as the (not quite standard) pair of *counterfactuals*:

- (Necessity) If the cause were to fail to obtain, the effect would not obtain (cf. Lewis's counterfactual account, though, as we'll see, with a different account of counterfactuals);
- (Sufficiency) If the cause, along with other relevant causes, were to obtain, the effect would obtain (cf. Mackie's INUS condition, though pruned of details).

To understand this, we need to say something about counterfactuals. A counterfactual is a conditional of the form 'If it had been P it would have been the case that Q'; it is standardly symbolized $P \Box \rightarrow Q$. Counterfactuals are very different in their meaning from the indicative conditionals that are much discussed in introductory logic courses. Consider the contrast between:

- (Indicative) If Oswald didn't shoot Kennedy, then someone else did. (did-did);
- (Counterfactual) If Oswald hadn't shot Kennedy, then someone else would have (had-would).

This isn't just a feature of English. Similar distinctions exist in other natural languages, although they are often marked by other linguistic devices (the use of the subjunctive, for instance). The indicative conditional might, or might not, be analysed using the material conditional of propositional logic; that is still controversial. The counterfactual certainly cannot be. It is clearly not truth functional.

The standard account of counterfactuals, proposed and developed independently by Lewis and by Robert Stalnaker, is, roughly:

$P \Box \rightarrow Q$ is true iff the consequent Q is true in the closest worlds in which the antecedent P is true

where closeness, though hard to characterize precisely, is a measure of similarity. So to evaluate a counterfactual with a false antecedent, you imagine the world that is most similar to the actual one but in which the antecedent is true, and then ask whether, in that world, the consequent is true.

A wrinkle: if, as seems plausible, the closest world to the actual world is the actual world itself, then if P and Q both actually obtain, it will seem that $P \Box \rightarrow Q$ will be trivially true: the closest world in which P is true (the actual world) is a world in which Q is true. Lewis accepts that conclusion (it is called 'strong centering'). But that means that (Sufficiency) is trivially true when the cause and the effect actually happen; it provides no constraint. If we want (Sufficiency) to be a contentful requirement, that is because we want to ensure that the causes would be sufficient to bring about the effect even if the world were somewhat different to the actual world. So, rather than following Lewis, we will need to read 'the closest worlds' to

include a *range* of worlds that are close to the actual world (that is why the counterfactuals are ‘not quite standard’).

People are sometimes suspicious of counterfactuals because they are vague and our evaluation of them may be sensitive to context (If Caesar had been in command in the invasion of Iraq would he have used atomic weapons? Or would he have used catapults?), and sometimes they seem to be indeterminate (If Verdi and Bizet had been compatriots would they have been Italian or French?). But vagueness, context sensitivity and indeterminacy are pervasive features of ordinary language. We use counterfactuals all the time, in science as well as in ordinary life (‘If we hadn’t added the catalyst, the reaction wouldn’t have happened’); we can’t do without them; and in most cases it is perfectly clear whether they are true or false.

To see how the two conditions work to give an account of causation, consider a pigeon trained to pick exclusively at red circles (the example is from Yablo). It now obediently picks at a circle of a particular shade of scarlet. Does the circle being *scarlet* cause the pigeon to peck? No, that’s too specific, it doesn’t meet the necessity condition. Does the circle being *coloured* cause the pigeon to peck? No, that’s not specific enough, it doesn’t meet the sufficiency condition. What causes the pigeon to peck is the circle being *red*. That meets both the necessity and the sufficiency condition. Causal claims need the right *grain*. Note then that causal claims are sensitive to the *properties* involved. So if we want to retain the idea that causation holds between states or events, those will have to be individuated in a way that reflects this: the event of painting something scarlet will be different to the event of painting it red, even though they happen at the same time. (Alternatively we could keep the idea that these are the same event, but deny that causation holds between events.)

Now we can state the second central idea: the neuro-states and the psychological states have their place in causal accounts that have different grains; so they give rise to different causal patterns at the different levels. this is the essence of List’s approach.

How can that happen? (We are not showing that it *must* happen; only that it *can*. The challenge was that it couldn’t.) Accept that along the bottom level of neuro-states, there is straightforward causation. But at the psychological level, that does not have to be replicated, even though there is supervenience. There are many different neuro-states that can underlie the same psychological states (multiple realizability); some would cause a further neuro-state that underlies one decision, others a further neuro-state that underlies a different decision. So if we go to close worlds in which the same psychological states obtain, and there will be radically different neuro-states underlying them, which cause neuro-states that underlie different choices. (Crucially, closeness at the psychological level is not the same as closeness at the neuro-level.) So the prior psychological states (beliefs and desires) don’t cause the decision, since whilst they meet the necessity condition, they violate the sufficiency condition (like having a merely coloured circle in the pigeon case).

But decisions, in contrast, can have, across all the close worlds, underlying neuro-states that are similar; so these in turn cause similar neuro-states, all of which underlie the same action (decisions have evolved to be modally robust in this way). If that is so, the decisions will meet both sufficiency and necessity with respect to action; so they count as causes of action.