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 grandfather. Could he succeed? We know he *doesn't*, since his grandfather lived 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; no relative to others: he is there. 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, relative to the full set of facts it was determined not to. Could you have stayed in bed this morning till 11 o'clock? Yes, relative to one set of facts (no one was forcing you up); no relative to another (you got here on time). 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 true. 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. 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* (desires, 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.

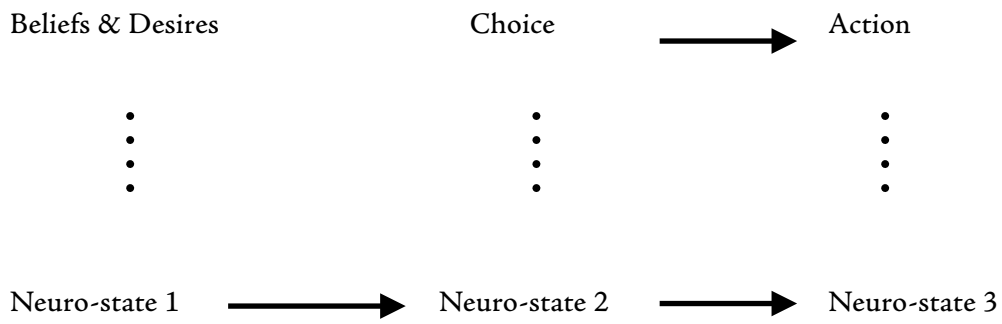
METAPHYSICS

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. 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 a screen supervenes on the arrangement of the pixels). The we can ask whether there is a consistent account in which the following desiderata obtain:

(i) at the psychological level: prior psychological states (beliefs and desires) do not cause action; 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 transistivity of causation, neuro-states on which prior psychological states (beliefs and desires) supervene cause neuro-stateson which action supervenes.

Or 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 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 is a measure of similarity. So to evaluate a counterfactual you imagine the world that is most similar to the actual one in which the antecedent is true, and then ask whether, in that world, the consequent is true.

A wrinkle: if 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 have happened; 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 we will need to read 'the closest worlds' to include a range of worlds that are close to the actual worlds (that is why the counterfacuals are 'not quite standard').

Counterfactuals might be vague and our evaluation of them may be sensitive to context (If Caesar had been in command in Korea would he have used the atom bomb? Or would he have used catapults?), and sometimes they may be indeterminate (If Verdi and Bizet had been compatriots would they have been Italian or French?); but then almost all of our language is like that. 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'), and in many cases it is perfectly clear whether they are true or false.

Now we are in a position to think about whether there could be causation at the neuro-level, but not at the psychological. To warm up, consider a pigeon trained to pick exclusively at red circles (the example is from Yablo). 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. Does the circle being *red* cause the pigeon to peck? Yes, that meets necessity and sufficiency. Causal claims need the right *grain*.

Now we can state the second central idea: the neuro-states and the psychological states fit into causal accounts that have different grains, so they give rise to different causal truths.

How can that happen? (We are not showing that it must; only that it is possible. The challenge was that it couldn't be.) 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 realisability'); some would cause a further neuro-state that underlies one decision, others a further neuro-state that underlies a different decision. Go to close worlds in which the same psychological states obtain, and there will be radically different neuro-states underlying them. (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. If that is so, the decisions will meet both sufficiency and necessity with respect to action; so they count as causes of action.