**Why are Causes the Right Way to Produce Effects?**

**David Papineau**

*Physics, Time and Agency Conference*

*Dublin*

*Irish Society for the Philosophy of Time*

*28-29 June 2022*

**A Background**

I’ve elsewhere defended a sort of regularity theory of causation with built-in causal asymmetry. It is motivated by the success of *causal inference* techniques that derive causal conclusions from patterns of unconditional and conditional correlations.

“The Causal Structure of Reality” 2021 *PhilSci-Archive* 1-51;“The Statistical Nature of Causation” 2022 *The Monist* 105:247–75.

Causally related variables enter into systems of (pseudo-)deterministic laws (equations) ordered by the requirement that a dependent (l.h.s.) term can only appear as an independent (r.h.s.) term after it has appeared on the l.h.s.

*Illustration:*

X 🡨 ex

Y 🡨 aX + ey

Z 🡨 bX + cY + ez

We can generalise to any variables X1, . . . Xn, and exogenous terms (r.h.s. only), e1, . . . en and any recursive laws (variables might be dichotomous, or determinable, or ordinal, as well as real-valued, and laws not just linear):

 Xi 🡨 F(X1, . . . Xi-1, ei)

*Theory*: Such an ordering in a system of laws will match causal ordering **iff** the *exogenous* e-terms are *probabilistically independent* of each other.

*Rationale*: The causal inference techniques can only be understood if we suppose that causal dependence is so constituted.

**B Actual Causation and Counterfactuals**

These laws (RLIs) are by their nature generic. How do they relate to *single-case* causation and counterfactual dependence? Here I appeal to the wealth of work in the Lewisian tradition that has devised recipes for reading answers to these questions off from “causal models” that specify how dependent variables are deterministic functions of others.

ST SH

BT BH S

What do the arrows mean here? Existing work takes them to portray either some primitive causal dependence or complex counterfactuals. I can do better: the models answer to how the variables are related in an RLI.

This opens the way to explaining counterfactuals (and actual causation) in terms of RLIs. *Interventionist semantics* for counterfactuals: set the relevant variable to the counterfactual value, leave all non-descendants unchanged, crank the equations . . .

**C Rational Action**

When is it rational to do A in pursuit of B?

I say: when it is probable that B “depends causal-counterfactually” on A. (Note how this explains rational action in terms of counterfactuals in terms of causation in terms of RLIs . . . I’m going to say “causal-counterfactually” to remind us I’m analysing counterfactuals in terms of causation.)

The “probable” here is non-eliminable. *Smoking causes cancer* and so it’s rational not to smoke, but *my avoiding cancer* mightn’t causal-counterfactually depend on *my not smoking*, and I’ll never know enough to know for sure whether it does or not. The causal difference smoking makes to cancer is the *probability* that you’re in a situation where cancer-avoidance causal-counterfactually depends on non-smoking.

The “depends causal-counterfactually” is also non-eliminable. It’s not enough that Pr(B|A) > Pr(B). Accidents are fewer among those who buy good car insurance, but that doesn’t make it rational to buy insurance to avoid accidents—because the correlation *doesn’t* evidence any single-case causal-counterfactual dependence of avoiding accidents on buying insurance (as is shown by the way the correlation disappears when we control for the factor on which accident-avoidance *does* causal-counterfactually depend, viz. a cautious disposition).

What about the fact the smoking-cancer correlation disappears if we control for some factor causally intermediate between smoking and cancer, such as fractured lung-cell DNA? But this doesn’t show cancer-avoidance is never causal-counterfactually dependent on non-smoking, just that it never *directly* so depends.

**D Agency Theories**

Agency theorists will say this is all the wrong way round. I am explaining rational action in terms of an independent counterfactual metaphysical dependency, the asymmetry of which is grounded in the asymmetry of causation which is grounded in the asymmetry of RLIs. Wouldn’t we do better to explain causation and its asymmetry in terms of *when it’s rational to act*?

After all, my account doesn’t do anything to explain *why* rational action is rational. What’s so good about doing A in pursuit of B just in case . . . [some complex story involving RLIs]? Shouldn’t we rather start with the simple idea that it’s rational to do A in pursuit of B just in case doing A renders B likely, and then—hopefully—explain causation in terms of that?

Newcombe’s paradox drives the point home. What’s so good about acting on causes? If you’re so smart, why aintcha rich?

But of course Newcombe’s paradox is a double-edged sword for agency theories. Buying good car insurance renders accident-avoidance likely . . .

The next two sections will argue against agency theories. I’ll then turn to Albert and Loewer, who at first pass might seem to have much in common with agency theories. I’ll show however, that they are much closer to my line. Having shown this, I’ll argue that my line is preferable to theirs.

**E Freedom**

There are two lines of agency responses to “medical” Newcombe cases.

Some (Glymour, Hitchcock? Woodward? Price?) pursue a bad line of thought: rational agents choose *freely*; so *their* decisions won’t be correlated with the other causes (cautious disposition) of desired outcomes (few accidents); so they can conclude that their action A won’t render B likely; and so they can avoid acting on the spurious A-B correlation.

Hm. We can debate about exactly what free action requires . . . but it surely *won’t* require that my cautious disposition doesn’t make me more likely to buy expensive insurance than less prudent people. (Of course, in making a decision, we can *compute* the A-B influence *on the pretence* that A is decorrelated from the other causes of B—ie on the supposition that we’re super-free. But that’s not because we really are super-free, but just because this is a good way to work out how likely it is B is *counterfactually* dependent on A—which is zero in this case, since good insurance adds no cases to those in which accident-avoidance was already counterfactually dependent on caution, as I observed earlier.)

**F Tickles**

So good agency theorists take a different line—the tickle defence. They start from the good point that decision-theoretic reasoning should always work within the reference class defined by everything you *know* about your situation (cf principle of total evidence). And then they argue that you will always know something (eg that you’re cautious) that will render any non-effective A probabilistically irrelevant to B (among cautious people expensive insurance isn’t associated with fewer accidents).

I don’t think the tickle line works. Apart from placing unreasonable demands on conscious introspection, it also struggles with incontinent agents, whose actions are partly influenced by their decisions but also partly influenced by *subconscious* factors spuriously associated with the effect (cf Lewis). In such cases, there’ll still be an A-B correlation in the reference class fixed by agents’ knowledge of their motives, but it will clearly be irrational for them to be moved by this in their decision-making.

Tickle defenders typically respond by saying they want to focus on agents that are fully continent. Maybe that is reasonable if they are just aiming to analyse the *concept* of causation (but isn’t *continent* a causal notion? and anyway doesn’t the concept have many different sources?). But I don’t think this will do if we are after a metaphysical relation that’s self-evidently good to act on—after all, *correlated in the reference class defined by the agent’s knowledge* just isn’t good for x agents to act on.

So I think we must explain rational action in terms of causation, rather than vice versa. It’s not rational to buy expensive insurance to prevent accidents (when you know the correlation is spurious) simply because the spuriousness of the correlation tells you that accident-avoidance is never causal-counterfactually dependent on buying insurance. And you can know this without attending to tickles and placing yourself in some narrow reference class.

Note how buying insurance can provide non-introspecters with new evidence that they’re cautious, and so that they’ll avoid accidents. Buying insurance has placed them in a reference class in which it’s likely they’ll avoid accidents. But this doesn’t make it rational for them to buy insurance. Similarly, putting yourself in a reference class that makes it likely the predictor’s put a million in the opaque box doesn’t make it rational to one-box, if the million’s presence isn’t causal-counterfactually dependent on your one-boxing.

**G Albert and Loewer**

David Albert and Barry Loewer have views which at first sight look like the agency line. But in truth they are much closer to me.

They don’t do counterfactuals in terms of causation, as I do, but go straight to counterfactuals (and aim to explain causation in terms of that).

They assume a metaphysics of deterministic statistical mechanics which they call “The Mentaculus”. Macrostates are equally likely to be in any of the microstates consistent with them. The universe starts in a low entropy macrostate (“The Past Hypothesis”) and then evolves into macrostates of increasing entropy.

Albert and Loewer say (very roughly) that B counterfactually depends on A if adding A to the current macrostate of the world (and supposing basic physics and the past hypothesis) makes it very likely that B, when subtracting A doesn’t.

They hold that many forward-facing counterfactuals (B after A) are true, but no (or few) backward-facing ones. This asymmetry derives from the asymmetry of (near) overdetermination: the present contains many macro-traces of the past, but none of the future (cf Lewis). So having A rather than not-A, plus present macro-facts, can make a big difference to the probability of future events but not (normally) to past events.

This line will also allow them to explain why it’s not rational to act on spurious correlations like that between buying good insurance and accident-avoidance. Since the earlier cautious disposition will have left many present macro-traces, the insurance won’t make any probabilistic difference in the reference class defined by those traces. Just as macro-records screen off past events from present actions, so do they screen of future events that are spuriously correlated with present actions.

This might look like Albert and Loewer are like agency theorists, getting counterfactuals (and causation) simply by considering which antecedents/actions make results B likely in the relevant reference classes.

But not so. The crucial point here is that their reference classes aren’t defined by *agents’ knowledge*, but by *all present macro-facts*. Many of these macro-facts might well be unknown to agents. After all, why should the present traces of my past caution necessarily be epistemically available to me? And Albert and Loewer certainly don’t argue that they will be—since they don’t need to, in order to hold that insurance isn’t correlated with accident-avoidance *given all current macro-facts*.

So Albert and Loewer come out like me here. Rational agency is explained in terms of actions making a counterfactual difference, and this is explained in terms of objective metaphysical patterns. Actions which remain correlated with putative results given only agents’ knowledge might not ever make such a counterfactual difference, if the correlation is spurious, and then both they and I say it won’t be rational to perform them.

**H Albert and Loewer vs Papineau**

The difference between Albert and Loewer and me thus lies in how we do counterfactuals. I do it in terms of asymmetric RLIs and the resulting recipe for reading off causal-counterfactuals. They do it more simply in terms of probabilistic differences given current macro-facts.

Albert and Loewer might seem to have an advantage over me here, resulting from the similarity they do bear to agency theories. Go back to the complaint that my account doesn’t do anything to explain *why* rational action is rational. As I asked, what’s so good about doing A in pursuit of B just in case . . . [some complex story involving RLIs]? Albert and Loewer look better off on this score. Isn’t there something intuitive about the idea that you should perform actions that make desired results likely in the reference class *defined by present macro-facts*?

Well, maybe. But if that’s so, I can appeal to the intuition too. This is because the asymmetry of (near-) overdetermination is a corollary of my own RLI account of causation. The various effects of a given cause will inevitably be correlated, given the independence of the exogenous terms. Which is to say that the many different effects will tend to co-occur whenever the cause does. So I too can say that it’s a good idea to act on causes because then you’ll generally be making desired results likely given present macro-facts.

**I No-Trace Cases**

Still, is it self-evident that it’s a good idea to perform actions that make desired results likely in the reference class *defined by present macro-facts*? Where did that come from? Why specifically *present* and *macro-facts*? Let me finish by putting pressure on this idea.

As is familiar, Albert and Loewer face difficulties in connection with those (overwhelmingly unlikely but naturally possible) cases where past events leave no present traces. They are forced to bite the bullet and say that in such cases there can be counterfactual dependence of past macro-facts on present events, and they seek to lessen the counter-intuitiveness by observing that in such cases agents will be in no position to exploit this in action.

Note than on my account no causal-counterfactual dependence obtains in such cases. I’ve just observed that on my account the inevitable correlation between common effects will mean that many of them will tend to co-occur in any given case. But this is a probabilistic phenomenon, and nothing rules out that all the common effects might sometimes be absent. Still, my asymmetric lawlike structures in the form of RLIs will still cover such cases and imply causal-counterfactual dependencies accordingly. So, on my account, Atlantis’s existence won’t causal-counterfactually depend on my not clicking my fingers, nor my accident-avoidance on my buying insurance, even if Atlantis or my caution leave no traces to screen off these results from these actions.

So perhaps we shouldn’t have been so quick to accept the idea we should perform actions that make desired results likely in the reference class *defined by present macro-facts*. The issue is clearest with forward-facing spurious correlations. Focus on the case where my caution happens to leave no present traces. Albert and Loewer will be forced to say my accident-avoidance now counterfactually depends on my insurance-buying, and will seek somehow to explain this away (presumably by reference to the unknowability of such unlikely counterfactual dependencies). But why not join me and simply say there’s no causal-counterfactual dependence, because of the structure of the underlying RLI laws?

So the truth, I’d say, is *not* that we should perform actions which make desired results likely given present macro-facts (and still less those which make them likely given agents’ knowledge), but rather actions on which desired results probably counterfactually depend in *my specifically causal sense*. (Such actions will still make desired results likely, but in a very specific sense—they make them likely within reference classes defined by the presence and absences of all the other causes of the desired effect.) We need to recognise that we are causal beings who have evolved causally in a causally structured world, and stop trying to analyse rational action without attending to this.