9

Causal and Explanatory Autonomy: Comments on Menzies and List

Ausonio Marras and Juhani Yli-Vakkuri

The chapter by Menzies and List offers a refreshing and much needed naturalistic perspective on the mental causation debate, and the debate over causation in the domains of the special sciences more generally. Though the literature they are responding to is concerned mainly with mental causation, Menzies and List are surely right that the answers they propose, if correct, generalize to any of the other presumably causal processes studied in the special sciences. By calling Menzies and List's perspective 'naturalistic', we mean to say that it represents the kind of philosophy that is sensitive to empirical science, its results and practices. It is a puzzling fact that 'naturalism' in this sense, though it has influenced—and, we think, improved—just about every other area of philosophy, has been so late in coming to the mental causation literature. But now we are confronted nearly with an embarrassment of riches: Raatikainen (2006), Shapiro and Sober (2007), the present chapter, as well as Menzies (2008), and Woodward (2008) all apply some version of the interventionist theory of causation to the problem of mental causation—a theory that has as good a claim as any to being an accurate account of the concept of causation at work in empirical science. These authors reach some interesting conclusions, some of which we think have a pretty good chance of also being true.

What's true, without a doubt, is that there is mental causation: mental events sometimes cause (mental and) other events, and even do so in virtue of being instances of mental properties. The contemporary *problem* of mental causation is that there are arguments, the most famous ones being due to Jaegwon Kim, that purport to show that if non-reductive physicalism is true, then there is no mental causation because—the reasons given for this vary—it cannot be that both mental events (or properties) and their physical realizers are causally efficacious, and our physicalist commitments somehow tell us that the physical realizers are the efficacious ones. Let us call these 'causal exclusion arguments'. The solution to the problem is to figure out what's wrong with the causal exclusion arguments.

The answer is, no doubt: *a lot*. Anyone who feels like challenging their ancillary premises—metaphysical platitudes, according to their advocates—has a lot to choose from. In our paper (2008), we laid out what we thought were the *non*-platitudinous assumptions in Kim's (2005) causal exclusion argument and singled out for criticism one that pertained to the identity conditions of events. However, this piecemeal approach of find-a-weak-premise-and-explain-why-it's-weak has never been successful at winning new converts to non-reductive physicalism, and in the concluding 'polemical remarks' in our paper (§10) we expressed some dissatisfaction with the way the entire mental causation game was being played: the concept of causation deployed by the disputants seemed utterly divorced from that at work either in common sense or science. A philosopher with a broadly naturalist outlook should show some interest—a lot of interest, in fact—in what scientists have to say about causation. When one does turn one's attention to scientific practice, we suggested, one will find weightier reasons to doubt the reality of *physical* causation than of mental or other higher-level causation: it is only in the special sciences, after all, that scientists explicitly take themselves to be investigating causal relationships (does smoking cause lung cancer?, etc.). Even a principle as supposedly fundamental as the causal closure of the physical domain, then, could not be taken for granted on a naturalistic approach. In n. 40 (Marras and Yli-Vakkuri 2008: 129) we suggested, however, that some form of the interventionist theory of causation could be used to vindicate the idea that there is causation going on, even at the 'bottom' level of fundamental physics. Indeed, it now seems to us that the correct interventionist theory of causation, supposing it is something along the lines of Woodward (2003), would, jointly with the assumption that mental and other higher-level properties are realized by physical properties, imply that every event that has a cause at all has a simultaneous physical cause—a principle even stronger than that which Kim (2005: 43) calls 'Closure'. Others, however, have reached different conclusions on the basis of different versions of interventionism.

In their contribution, Menzies and List seek to use a version of the interventionist theory to redefine the assumptions of the mental causation debate. While we think this is just what the debate needs, we find some of the conclusions Menzies and List reach questionable. In the following we will try to articulate our main concerns about their argument for, as well as attribution of, the 'causal autonomy' thesis, and the responses to the causal exclusion arguments that are implicit in their contribution.

NON-REDUCTIVE PHYSICALISM AND CAUSAL AUTONOMY?

Menzies and List attribute to non-reductive physicalists the view that the systems studied in the special sciences 'have causal powers that are independent of

those of their more basic physical properties' (this volume: 108). No doubt this claim is true of the British emergentists of the early twentieth century, but is it true of contemporary non-reductive physicalists? Menzies and List cite Jerry Fodor as an example of a contemporary non-reductive physicalist who holds this view. But to attribute this view to Fodor is, to say the least, surprising: didn't Fodor (1985: 42) famously claim that 'if mind/brain supervenience goes, the intelligibility of mental causation goes with it'? The point of Fodor's remark was that the systems studied by psychology, and, by extension, the special sciences in general, would have no causal powers *at all* unless their properties supervened on the underlying physical properties of their constituents. For Fodor, any non-reductive physicalist who does not believe in magic must accept that the causal powers of special science systems are *dependent* on, and *determined* by, the causal powers of their more basic properties. No doubt every non-reductive physicalist must insist that mental—or, in general, higher-level—properties are *distinct* from the lower-level properties that realize them, and that, consequently, their causal powers must also be distinct (assuming, as most now do, that 'real' properties are individuated by their causal powers); but to assert the *distinctness* of the former properties and causal powers from the latter properties and causal powers is not the same as to assert their *independence*. When Fodor argued for the *autonomy* of the special sciences he was arguing for their *explanatory* autonomy, not for the *causal* autonomy of the systems which they study. The main concern of Fodor (1974)—the *locus classicus* of Fodor's exposition and defence of the 'autonomy thesis'—was to deny the *reducibility* of the special sciences in a sense of 'reduction' that would require the coextension of each special science predicate with a predicate of physics, while insisting that the real and legitimate aim of reduction should be to 'explicate the physical mechanisms whereby events conform to the laws of the special sciences' (1974: 27). Quite clearly, the point of explicating such physical mechanisms is precisely to show *how* special science systems and their properties manage to be causally efficacious: they do so by way of the physical mechanisms which implement them. To suppose otherwise would have been, from Fodor's point of view, to believe in occult powers.

Now it is a separate question whether Fodor was *right* to link the possibility of mental causation with psychophysical supervenience as he did—a matter to which we will return. However, the conception of higher-level causation as 'working through' underlying *physical* causal processes, which is presupposed in Fodor (1974), is widely shared by physicalists, of both the reductivist and non-reductivist variety. That being the case, the basis for attributing to the latter the view that the causal powers of higher-level properties are independent of those of physical properties is not clear.

But what do Menzies and List mean by saying that 'the higher-level properties of . . . systems [studied in the special sciences] have causal powers that are independent of those of their more basic physical properties'? The sense in which

they think this is true turns out to be a bit surprising. While one might have thought that an affirmation of an 'independence' thesis is equivalent to the denial of a supervenience thesis (so that two systems might differ with respect to the causal powers of their higher-level properties while being indiscernible with respect to the causal powers of their physical properties), Menzies and List, apparently, simply mean to say that some higher-level properties have causal powers which are not causal powers of any of their physical realizers. Though more modest than one might have expected, the thesis is certainly interesting. The examples by which they attempt to establish this thesis involve higher-level properties whose physical realizers are 'too specific' to count as causes of their effects. In Menzies and List's argument, both a higher-level property and its physical realizer may turn out to be, so to speak, 'causally excluded'—which is excluded depends on contingent, empirical facts about each case.

This argument is based on their version of the interventionist theory of causation. In effect, Menzies and List use interventionist ideas to motivate a principle that is very nearly equivalent (see note 1) to the 'proportionality constraint' of Yablo (1992). The former does seem like an improvement on the latter in at least one respect: while Yablo's proposal was deeply involved in—some might say marred by—intuition-driven essentialist metaphysics, Menzies and List's proposal is presented as being in accord with good scientific practice. Nonetheless, the outcome is very similar, and it bears asking whether the outcome is true.

It is worth pausing to consider just why Yablo's and their proposals assign the same truth values to the causal judgements Menzies and List consider. They do so for very nearly the same reason. Consider, for example, the two rival judgements about Yablo's (1992: 258) pigeon example:

(Red) The triangle's being red caused the pigeon to peck.

(Crimson) The triangle's being crimson caused the pigeon to peck.

Given that the pigeon had been trained to peck at red things, (Red) satisfies Yablo's proportionality constraint but (Crimson) does not. (Red) satisfies the constraint because both of the counterfactuals, 'Had the triangle been red, the pigeon would have pecked' and 'Had the triangle not been red, the pigeon would not have pecked', are true, whereas (Crimson) does not satisfy it because only the first of the two counterfactuals, 'Had the triangle been crimson, the pigeon would have pecked' and 'Had the triangle not been crimson, the pigeon would not have pecked', is true. The latter is false because in one of the nearest possible worlds in which the triangle is not crimson, the triangle is some other shade of red, resulting in the pigeon pecking. So on Yablo's proportionality account (Red) is assigned True and (Crimson) False; i.e. the triangle's being red, not crimson, was the cause of the pigeon's pecking. But the very same truth value assignments, and essentially for the same reasons, would fall out of Menzies and List's account; the only difference is that the antecedents of the relevant counterfactuals are

understood as being made true by an intervention.[1] The reason (Crimson) would be false, on their view, is that there is an intervention I—namely one that changes the triangle from crimson to another shade of red—such that if the triangle had been non-crimson as a result of I, the pigeon would have pecked regardless.

Applying similar reasoning to claims that ascribe causal efficacy to higher-level properties (instead of redness) versus their physical realizers (instead of crimson) yields the desired conclusion. Given Menzies and List's version of the interventionist theory of causation, it is overwhelmingly likely that there are *some* cases in which the higher-level property but not the realizer will turn out to be a cause—but this will be, as they correctly point out, an empirical question.

Is this reasoning sound? We have our doubts. The weakest part of the case for the 'causal autonomy' thesis is Menzies and List's version of the interventionist theory, which they present as a 'simplified' version of Woodward's. Let us consider the differences between the two theories. Menzies and List are (roughly[2]) committed to the following.

(ML) A causes (or, as Menzies and List say 'makes a difference to') B iff (i) $A \mathrel{\Box\!\!\rightarrow} B$; and (ii) $\sim A \mathrel{\Box\!\!\rightarrow} \sim B$.

These are 'interventionist counterfactuals', in which the antecedent is to be understood as being made true by an intervention in an ideal experiment. This stipulation invalidates the rule $A, B/(A \mathrel{\Box\!\!\rightarrow} B)$, which is valid in Lewis's (1973) counterfactual logic since, clearly, the bare truth of A and B does not guarantee that B would still be true if A were made true in an ideal experiment. However, apart from a minor technical revision made to accommodate this fact, Menzies and List's counterfactuals are understood as having the familiar Lewisian semantics. It follows, then, that if $A \mathrel{\Box\!\!\rightarrow} B$ is true in a world w, then B is true in *every* world w' that resembles w as much as A's being made true by an intervention in w' will allow.

[1] This, of course, makes a difference to truth conditions, so Yablo's and Menzies and List's proposals do not assign exactly the same truth conditions to causal claims (for example, it appears that on Yablo's proposal a barometer reading might qualify as a cause of a storm). However the truth *values* they assign in the cases that interest us are the same.

[2] This is essentially a notational variant of Menzies and List's account. To take into account the 'contrastive' character that causal claims have according to interventionism, they use a more cumbersome notation which quantifies over (possibly many-valued) variables and values assigned to them. Since they themselves focus on simple cases involving binary variables representing the occurrence or non-occurrence of an event, and since in nearly all discussions of mental causation the examples concern causal relations obtaining between *token* events (e.g., Jones's desire for water causes him to reach for the glass), which are naturally reported using sentence nominalizations ('Jones desiring water', 'Jones reaching for the glass') that take one of two values, we will for simplicity's sake mostly use 'A' and '∼A' in place of 'A = 1', 'A = 0', and the like. Sometimes, however, it will be more natural to speak of a binary variable 'A' 'changing' its value, so we will alternate between the two idioms. Notably, (W) below is stated using the idiom of variables and values.

Woodward, on the other hand, is (roughly) committed to this.

(W) *X* causes *Y* iff [there is an intervention *I* on *X* such that if *I* were to change the value of *X*, then the value of *Y* would also change].

This too is a simplification,[3] but it preserves an essential feature of Woodward's theory which is not present in Menzies and List's: namely, that the right-hand side of (W) has the form of an *existential generalization* over interventions. The right-hand side of (ML), on the other hand, is a conjunction of two claims which are, *in effect*, universal generalizations over interventions.

Here's what we mean by the 'in effect'. Let us abbreviate '*A* is made true by an intervention' as '**I**(*A*)'. Since Menzies and List assume, *mutatis mutandis*, the usual Lewis semantics, the right-hand side of (ML) is true iff [*every nearest* **I**(*A*)*-world is a B-world* and *every nearest* **I**(∼*A*)*-world is a* ∼*B-world*]. In other words, a claim is being made about *all* interventions that bring about *A* (or ∼*A*) and which occur in worlds resembling the actual world as much as the truth of **I**(*A*) (or **I**(∼A)) will permit.

On Woodward's theory, on the other hand, the existence of even *one* intervention on *X* that would alter *Y* implies that *X* is a cause of *Y*. It is clear that (ML) and (W) deliver different verdicts about (Red) and (Crimson). According to (W), both (Red) and (Crimson) are true, since there is an intervention that 'changes' the redness of the triangle (i.e., makes it non-red) under which the pigeon's pecking would be 'changed' (i.e., the pigeon would not peck), as well as one that 'changes' the scarlet-ness of the triangle (making it non-scarlet by making it a non-red colour), under which the pecking would be 'changed'. One can reason similarly about mental properties and their physical realizers.

Woodward uses up a few pages explaining why 'Causal Claims [Tell] Us What Happens Under Some (Not All) Interventions' (2003: 65), and we will not repeat what he has to say here. Rather, we will say what seems to us correct about Menzies and List's version of the proportionality constraint. If we replaced the words 'causes' with 'causally explains' and the 'iff' with 'only if' in (ML), we would have, we think, a plausible claim—call the resulting claim (ML*). A good causal explanation obviously does more than make a true causal claim—for example, 'The cause of lung cancer causes lung cancer' may be true but explains nothing. 'Smoking causes lung cancer' is more enlightening but still leaves something to be desired: in particular, the latter claim does not tell us *which* interventions on smoking would affect lung cancer and how, but only that some would. Good causal explanations specify the relationship between two variables *X* and *Y*—say, by means of an equation $Y = F(X)$—in a way that tells us just how *Y* would change under interventions on *X*. When what is to be explained is a single token event, which can be viewed as the taking of a particular value (1 or 0) by a binary variable according to whether the event occurs (1) or doesn't occur (0),

[3] The full account is given in Woodward (2003: 59), where it is labelled '(**M**)'.

it seems plausible that the explanation should specify a variable by means of the manipulation of which the event could be made to both occur and not occur. If the explaining variable is also binary, then the right-hand side of (ML*) seems a plausible condition for adequacy of the explanation. (But (ML*) would only state a necessary condition, as it does not rule out claims like 'That the cause of the fire occurred causally explains why the fire occurred'.)

It is an important point, made by Batterman (2002) and Woodward (2003: 231f), among others, that often lower-level explanations of phenomena are simply inappropriate in science—for example, explanations that cite the positions of each of the 9×10^{70} molecules that compose a thermodynamic system do not adequately answer questions like 'Why is the pressure of the gas *P*?' This is arguably, as Woodward does argue, because they do not provide us with information that we could use for manipulation of the explanandum (moving the individual molecules around is not a very good strategy for altering the pressure of the gas). This is another sense, besides Fodor's, in which the special sciences have *explanatory* autonomy, but the case for the *causal* autonomy of the systems they study remains to be made.

CAUSAL EXCLUSION ARGUMENTS

What is the reply to the causal exclusion arguments, in particular to Kim's (2005: ch. 2), that is implicit in Menzies and List's chapter? It is evident that this would entail rejecting Closure—the assumption that, according to Kim, guarantees the result that the physical cause will 'win' whenever mental properties and their physical realizers 'compete' for causal efficacy. The very same considerations that militate in favour of Menzies and List's 'causal autonomy' thesis, *if* their (ML) is assumed, will militate against Closure. That an event *E* has a mental cause *M* occurring at *t* is no guarantee that it will have a physical cause also occurring at *t*, for any putative physical cause *P* occurring at *t* may be ruled out by (ML) as 'too specific', i.e., *P* may fail to satisfy (ML)(ii). Both Raatikainen (2006) and Menzies (2008) respond to the causal exclusion arguments on the basis of a similar understanding of the interventionist theory of causation.

There are two problems with this line of response. The first is that, as we have argued (Marras and Yli-Vakkuri 2008: 111), Closure is redundant to Kim's causal exclusion argument. Kim can make his case that non-reductive physicalism implies epiphenomenalism without that assumption (if he can make it at all). Menzies and List might, however, raise another objection to Kim's argument: two of Kim's implicit assumptions concerning how supervenience relates to causation—labelled 'SC I' and 'SC II' in our paper (2008: 106f)—turn out to be no more tenable than Closure, if (ML) is assumed. SC I says that an event *C* can only cause a (higher-level) event *E* by causing *E*'s supervenience base, and

SC II says that a (higher-level) event C can cause E only if C's supervenience base causes E. Given the proportionality constraint encoded in (ML), however, we have no reason to expect this to be the case: a higher-level event that causes E may be too non-specific to qualify as a cause of E's supervenience base, and the supervenience base of a higher-level event that causes E may be too specific to qualify as a cause of E.

The second problem is that, again, (ML) itself looks untenable, and if (ML) is false, the objections to Closure, SC I, and SC II we just sketched are unsound. If, as we suppose, (W), not (ML), is a (more nearly) correct account of causation, we can, in fact, give arguments for all of Closure, SC I, and SC II. To illustrate with Closure: suppose a higher-level event H causes another event E; then by (W) there is an intervention I that sets $H = 0$ such that if I were carried out, it would be the case that $E = 0$. Supposing H is realized by a physical event P_i, and that $P_1, \ldots, P_k$ are all the possible realizers of M, then there is an intervention on P_i that would set $E = 0$, namely one that sets $P_j = 0$ for each $1 \leq j \leq k$. (Why is it guaranteed that there is such an intervention? Because the intervention I that sets $H = 0$ itself is such that it sets $P_j = 0$ for each $1 \leq j \leq k$.) It follows by (W) that P_i also causes E. So, if an event has a higher-level cause, then it has a physical cause—this principle is, in fact, stronger than Kim's Closure.

What, then, is wrong with the causal exclusion arguments? We suggest that, if (W) is correct, the culprit is the causal exclusion principle itself, which is, in one form or another, common to all the arguments: in Kim's (2005: 42) version, it is the principle that 'No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of overdetermination'. Again, it seems plausible that both a higher-level event H and its physical realizer P can be intervened on in such a way as to bring about the non-occurrence of some putative effect E of H, showing both H and P to be causes of E.[4]

CONCLUDING REMARKS

Menzies and List are surely right that whether special science properties are causally autonomous or not is an empirical matter, and likewise the question of whether special science properties, if distinct from physical properties, are ever causally efficacious. We also agree with them that the interventionist theory, broadly construed, is a promising framework for answering these questions. However, within this framework, metaphysical questions concerning the truth

[4] We assume here Kim's 'fine-grained' conception of events, on which each event is an instance of exactly one property. If this assumption is not made, a different reply, which we outline in 'The "Supervenience Argument" ' (Marras and Yli-Vakkuri 2008), is available.

of causal claims, and epistemological questions concerning the adequacy of explanations, can and must be kept apart, and it seems to us that Menzies and List's attempts to both defend the causal efficacy of special science properties and argue for their causal autonomy founder on their conflation of these two types of questions. Mental and other higher-level causation is no less defensible for that, but there is, as far as we can see, no case for the causal autonomy of higher-level properties that does not rest on a conflation of explanatory adequacy with causal efficacy.

Finally, we would like to return to Fodor's claim that the possibility of mental causation depends on the truth of psychophysical supervenience. One surprising result that becomes evident as soon as we consider the question of mental causation within the interventionist framework is that Fodor was wrong about this. If interventionism is right, then mental causation is real just in virtue of the fact that there are relationships between mental and other properties that we can exploit for manipulation—nothing further is required. Fodor's conception of mental and other higher-level causation as 'working through' physical causal mechanisms is an *empirical hypothesis*; it is not an account of the *nature* of causation. On this we are, we presume, in agreement with Menzies and List, though we perhaps part company with them in tentatively accepting Fodor's hypothesis.

REFERENCES

Batterman, R. 2002. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press.

Fodor, J. 1974. 'Special Sciences, or the Disunity of Science as a Working Hypothesis'. *Synthese* 28: 97–115.

——1985. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.

Lewis, D. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Marras, A. and Yli-Vakkuri, J. 2008. 'The "Supervenience Argument": Kim's Challenge to Nonreductive Physicalism'. In F. Orilia and S. Gozzano (eds), *Tropes, Universals, and the Philosophy of Mind*. Frankfurt: Ontos Verlag.

Menzies, P. 2008. 'The Exclusion Problem, the Determination Relation, and Contrastive Causation'. In J. Howhy and J. Kallestrup (eds), *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*. Oxford: Oxford University Press.

Raatikainen, P. 2006. 'Mental Causation, Intervention, and Contrasts.' Unpublished. Available at www.mv.helsinki.fi/home/praatika/ (accessed 9 January 2008).

Shapiro, L. and Sober, E. 2007. 'Epiphenomenalism—the Do's and the Don't's'. In G. Wolters and P. Machamer (eds), *Studies in Causality: Historical and Contemporary*. Pittsburgh: University of Pittsburgh Press.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

——2008. 'Mental Causation and Neural Mechanisms'. In J. Howhy and J. Kallestrup (eds), *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*. Oxford: Oxford University Press, 218–62.

Yablo, S. 1992. 'Mental Causation'. *Philosophical Review* 101: 245–80.