THE EXCLUSION PROBLEM, THE DETERMINATION RELATION AND CONTRASTIVE CAUSATION

Peter Menzies
pmenzies@scmp.mq.edu.au

1. Introduction

Stephen Yablo's influential article "Mental Causation" made an interesting new move in the debate about the exclusion problem of mental causation. He observed that:

(a) determinables are not excluded from causal relevance by their determinates;
(b) the relation of mental properties to their underpinning neural properties is analogous to the relationship of determinables to determinates.

In this paper, I want to:

- Argue that Yablo is probably wrong about (b); and show that the account of causal relevance that he uses to motivate (a) is not satisfactory.
- Motivate and defend an account of causal relevance that does a better job than Yablo's of explaining the phenomena.
- Show how this account provides a better answer to one version of the exclusion problem.
- Show that this account yields a very surprising answer to a more recalcitrant version of the exclusion problem.

2. The Exclusion Problem and Yablo on Mental Causation

Here is one version of the famous exclusion problem:

(1) If an instance of a property F is causally sufficient for an instance of a property G, then no instance of a property F* distinct from F is causally relevant to the instance of G. (exclusion)
(2) For every instance of a physical property G, there is a physical property F such that an instance of it is causally sufficient for the instance of G. (causal closure of the physical)
(3) Every mental property F* is distinct from every physical property F. (property dualism)
(4) Therefore, no instance of a mental property F* is causally relevant to an instance of a physical property G.

In his paper, Yablo made two crucial observations.

(a) Determinables are not excluded from causal relevance by their determinates.
His example: A pigeon is trained to peck at red things to the exclusion of things of other colours: the pigeon is presented with a red triangle and she pecks it. He claims that the redness of the triangle was causally relevant to her pecking, even though the triangle was a specific shade of red, say crimson, which was causally sufficient for her pecking.

(b) The relationship between mental and physical properties resembles the relationship between determinables and determinates. He states the following necessary condition on the determination relation:

Property F *determines* G only if:
(i) necessarily, for all x, if x has F, then x has G; and
(ii) possibly, for some x, x has G but lacks F.

The relevant type of modality is metaphysical. If neural properties determine mental properties, then it is not surprising that the relationship between mental and physical properties has certain features: supervenience, and multiple realizability.

It follows from these observations that mental properties will not be excluded from causal relevance by their underlying neural properties. Like determinables and determinates, mental and physical properties do not compete for causal relevance.

However, Yablo also makes a stronger claim about the notion of causation. He imposes a proportionality constraint on the notion of 'a cause': a cause must be specific enough for its effect but no more specific than required. Let us say that a property instance Fa *screens off* a property instance Ga from a property instance Ha iff if it were the case that Fa but not Ga, then it would (still) be the case that Ha.

Then property instance Fa *is a cause* of property instance Ga iff
(i) If it were not the case that Fa then it would not be the case that Ga. (*contingency*)
(ii) If it were the case that Fa then it would be the case that Ga. (*adequacy*)
(iii) Fa screens off all instances of its determinates from Ga. (F is *enough* for G)
(iv) No instance of F's determinables screens Fa off from Ga. (F is *required* for G)

Yablo imposes condition (iii) to eliminate instances of properties that are not specific enough. Suppose a safety value stiffens because of structural defect so that its slow opening causes a boiler to explode: the valve's opening per se is not a proportional cause of the boiler's explosion because it doesn't satisfy (iii). He imposes (iv) to eliminate instances of properties that are too specific. Suppose Socrates can't drink the hemlock without guzzling: his guzzling the hemlock isn't a cause of his death because it doesn't satisfy condition (iv).

It is easy to see that the triangle's being red may be a cause of the pigeon's pecking. Likewise, a person's having a mental property may be a cause of some bodily movement.

3. The Determinable/Determinate Distinction

To evaluate Yablo's strategy we have to get a better understanding of the determinables/determinates relation. His account is not satisfactory as it doesn't rule out the property F determining the disjunctive property F v G, or the conjunctive property F & G determining the property F.

The best account of this relation I know is Eric Funkhouser's in his paper "The Determinable-Determinate Relation". The central insight of his account is that a determinate is a specification of its determinable along certain dimensions. He calls these the determination dimensions. For example, red can be specified along the three dimensions of hue, brightness, and saturation. Funkhouser proposes sufficient conditions for determination:

> Property F *determines* property G if (i) F and G have the same determination dimensions; and (ii) the range of determination dimension values for F is a proper subset of the range of determination values for G. Further, he says that a *superdeterminate* is a property that doesn't allow of further determination.

Let's understand the determination relation in terms of Funkhouser's framework. Then we might object to Yablo's strategy for dealing with the exclusion problem.

First, mental properties are not related to physical properties as determinables to determinates.

- The determination dimensions of mental properties seem to be different from those of neural properties. The determination dimensions for belief, for example, are content and degree of confidence. Any superdeterminate in the property space determined by these determination dimensions may be multiply realized by neural properties.

Secondly,the constraint of proportionality does not always lead to determinate results:

- Suppose that the pigeon has been trained to peck at redish things, including pink, and orange objects as well as red objects. But she has been trained not to peck at blue or green objects. She is presented with a red object and she pecks at it. The problem is it is hard to evaluate the counterfactuals (i) and (iv) in the definition of a proportional cause: if the triangle had not been red, would the pigeon have pecked? If the triangle had been coloured but not red, would she have pecked?

4. The Contrastive Nature of Causation

Yablo does not explain why causes must be proportional. I think it is possible to explain this in terms of the common dictum that causes must make a difference to their effects. This dictum implies that variation in the cause leads to variation in effect. I suggest that the best way to articulate this dictum is to reconstruct our causal judgements as judgements about relationships between variables: causal judgements tell us about how changes in the values of the causal variable are related to changes in the values of the effect variable. (Note that the relationship between a variable such as mass and a value such 10 grams is the relationship between determinable and determinate.)

A broad consensus has emerged among a group of philosophers of causation (Glymour, Spirtes, Scheines; Pearl; Woodward; Hitchcock) about how to capture the idea that a cause makes a difference to its effect within the framework of variables and values. It will help to make some simplifying assumptions: let us focus on deterministic systems that do not involve complicated processes of pre-emption and overdetermination.

There are two kinds of causation: causation between variables and causation between values of variables. (These more or less map onto the distinction between type- and token-causation).

(I) Type-causation (or causation between variables): X causes Y [relative to a kind of system S] iff if an intervention were to occur in a system of kind S to change the value of X, then the value of Y would change.

Some features of this account:

- An intervention on the variable X (with respect to variable Y)) is an idealised manipulation that sets the value of X independently of other possible causes of Y. This is a causal notion so that this is not intended to be a reductive account.
- If type-causal claims relate event-types, these claims must be translated into claims relating variables, usually binary variables.
- The interventionist account is often framed in terms of counterfactuals: e.g. there are distinct values of X (say, X = x, x' such that x ≠ x') and Y (Y = y, y' such that y≠y') and there is a system of kind S such the following two counterfactuals are true:

  If an intervention that sets X = x were to occur in the system, then Y =y;
  If an intervention that sets X = x' were to occur in the system, then Y =y'.

  In interventionist accounts of counterfactuals, the notion of an intervention plays the same role as the notion of a miracle in Lewis's similarity account of non-backtracking counterfactuals.
- On this account, a cause makes a difference to its effect in the sense that it is X taking value x rather than x' that causes Y to take the value y rather than y'. *Causal claims have an implicit contrastive structure.*

(II) *Token-causation (or between values of variables)*: X=x causes Y=y [relative to a particular system of kind S] iff there are values of X (say, x*) and Y (say, y*) such that if an intervention were to occur in the given system to change the value of X from x* to x, then Y would change from y* to y.

- Token-causal claims involving events or property instances can be translated into this framework by representing the occurrence/non-occurrence of the events or property instances in terms of variables taking certain values.
- Again this account can be expressed in counterfactual terms: e.g. there are distinct values of X (x, x* with x ≠ x*) and Y (Y = y, y* such that y≠y*) such the following two counterfactuals are true about the given system of kind S:

  Given that X = x* occurs in the given system, then Y =y*;
  If an intervention that sets X = x were to occur in the system, then Y =y.

- This account of token-causation focuses on a given actual system; and it anchors the changes in the values of the variables to certain baseline or default values, x* and y*. Sometimes these are made explicit in contrastive statements: it was the fact that X=x *rather than* X=x* that made the difference to the fact that Y =y *rather than* Y =y*. More often they are determined contextually. (See Hitchcock, Maslen, Schaffer) I claim that these default values are set in a systematic way: they represent the natural or normal or to-be-expected state of the system antecedent to any intervention.
- *Again this account makes token-causation essentially contrastive in character*. This crucial feature explains many aspects of our token-causal judgements.

(A) *Event aspects*
We explain Socrates' death by saying he drank *hemlock* at dusk, not by saying that he *drank* hemlock at dusk or drank hemlock *at dusk*. Some have invoked event aspects or event allomorphs to explain this. The contrastive account sees such examples as illustrating of contrastive focus.

(B) *Quantitative variables*.
Suppose that giving more than 100mg of a drug will cure a patient's disease. You can give the patient no dose, 100 mg, 200mg or 500 mg. You give him 200mg and he recovers. Did your giving him the 200 mg dose cause him to recover? The contrastive account says it depends on the default values of the variables. Lewis' counterfactual theory has trouble here because it says to excise completely the event of giving 200 mg dose in evaluating the counterfactual 'If you had not given the 200 mg dose…'.

(b) *Fragile events and the asymmetry between hasteners and delayers*.
Suppose a doctor gives a patient a large dose of morphine that hastens the patient's death. The doctor's action was a cause of the patient's death. Lewis says that we sometimes take events to be fragile in the sense that they have essential times and manners of occurrence. But now suppose that doctor delays a patient's death by resuscitating him for a short time. The doctor's action didn't cause the patient's death. How to explain this seemingly arbitrary asymmetry between hasteners and delayers? The examples are set up so as to cue us into thinking about the effect in terms of the contrast between dying-at-t and dying-at-later-time t'. There is no objective basis for treating hasteners and delayers differently in terms of this contrast. Now consider a different contrast: the patient's being-alive-at-t and dying-at-t. A hastener can make a difference with respect to this contrast in a way that a delayer cannot.

5. Application to Exclusion Problem

I agree with Yablo that his example about the pigeon demonstrates the falsity of the exclusion principle. But I think the contrastive account of causation provides a better explanation of our causal judgements about the example. Recall that we have the rival causal judgements:

> *Red*: The triangle's being red caused the pigeon to peck.
> *Crimson*: The triangle's being crimson caused the pigeon to peck.

In order to apply the contrastive account to the example, we have to determine the relevant variables and the relevant contrasts. Suppose that the variables and the contrasts are the obvious ones. So these causal statements come to:

> *Red*: The triangle's being red rather than not red made the difference to the pigeon's pecking rather than not pecking.
> *Crimson*: The triangle's being crimson rather than not crimson made the difference to the pigeon's pecking rather than not pecking.

The structure of the situation is illustrated in the Figure 1.

(The contrastive account also explains our judgments about the examples Yablo used to motivate the conditions for his principle of proportionality. The *enough* condition (iii) is supposed to ensure that the cause is specified in sufficient relevant detail. But this is captured by the contrastive nature of token-causation: Socrates' drinking hemlock rather than not drinking it made the difference with respect to his dying rather than living; his guzzling rather than sipping it didn't make the difference. The *required* condition (iv) is supposed to ensure that the cause is specified with no more detail than is necessary. Again, the contrastive account has the same effect: the valve's openly slowly rather than speedily made the difference to the boiler's exploding or not exploding; whereas the valve's opening rather than not opening did not.)

So the case of the pigeon demonstrates the falsity the exclusion principle. The same moral seems to apply in the case of mental causation. Consider a particular case in which we are considering whether an instance of a mental property M causes some physical behaviour that instantiates the property B. Let us suppose that the property M can be realized by several neural properties but on the given occasion M is realized by N_i. Let us suppose that Ni is causally sufficient for the instance of B. The contrastive theory implies that the causal sufficiency of Ni does not undercut the causal efficacy of M with respect to B. Consider two cases.

Case 1. Suppose that Yablo is correct in thinking that realization is the same as determination. Then the situation is structurally similar to that depicted in Figure 1. It is the contrast between M and not-M, and not the contrast between N_i and any other neural realiser of M, that makes the difference to the contrast between B and not-B.

Case 2. Suppose that realization is not the same as determination, which I think is likelier to be the case. Then the situation is depicted in Figure 2 above. The difference here is that while N_i realizes M on the occasion, the other values of the relevant neural variable need not be understood as realizers of M. It is probable that the contrast between B and not-B is to be explained in terms of the contrast between M and not-M, and not the contrast between N_i and some other value of the relevant neural variable.

6. Coda

The common version of the exclusion principle, formulated by Yablo, is false. This particular version is a hangover from a period when advocates of regularity theories of causation thought that causation could be understood in terms of nomologically sufficient conditions.

This raises the question: What happens if we reformulate the exclusion principle, replacing the reference to causal sufficiency with reference to causation proper, where this is understood in terms of the contrastive theory? In particular, consider this version:

> If an instance of property F causes an instance of property G, then no instance of property F* distinct from F causes the instance of G. (*exclusion reformulated*)

I think that this general principle is easily shown to be false: overdetermination turns out to be very common.

A special instance of this principle has been the focus of the debate about mental causation. The advocates of non-reductive physicalism have wanted to say that the following instances of the principle are false, where M is a mental property with possible neural realisers $N_1,\ldots, N_n$:

> (*Top-bottom*):If an instance of the property M causes an instance of a physical behavioural property B, then no instance of a neural realiser property N_i causes that instance of B
>
> (*Bottom-top*): If an instance of a neural realiser property N_i causes an instance of B, then no instance of M causes that instance of B.

Non-reductive physicalists want to deny these principles because they believe that mental properties have causal powers independent of those of their neural realizer properties. Can I say that these principles are false? I wasn't sure of the answer to this question when I noticed the following surprise result:

Supervenience of mental properties on physical + *contrastive theory of causation* ⇒ *type identity of mental and physical properties*

Let us focus on the situation in which an instance of M causes an instance of B, where M and not-M realized by properties $N_1,\ldots, N_4$. The situation is depicted in Figure 3.

This is to be interpreted thus:

- The inner square in the outer square represents the set of possible systems of the given kind that we are considering: human beings characterised in terms of both mental and neural variables and conforming to the laws of intentional psychology and the laws of neurology. Our counterfactuals are to be evaluated with respect to this set of possible systems.
- We have a partition of this subspace in terms of a mental variable M and each cell of the partition is subdivided into its possible neural realizers: N_1 and N_2 realize M and N_3, and N_4, realize not-M.
- The areas that are shaded represent the possible systems that exhibit an instance of the behavioural property B.
- The figure shows that M causes B, and also for any particular system a, Ma causes Ba.

It is easily seen that it follows that every system x, the following are true:

If it were N_1a then it would be Ba.
If it were N_2a then it would be Ba.
If it were N_3a then it would be not Ba.
If it were N_4a then it would be not Ba.

Suppose that N_1a is the actual realizer of Ma in the particular system a under consideration. These counterfactuals do not show by themselves that N_1a causes Ba. *The contrastive theory of causation says that we have to ask: Which variable is N_1a a value of and does variation in this variable lead to variation in the effect variable?*

To answer this question let us zoom out and get a broader perspective on the situation. We see that the counterfactual dependences between M and B are mirrored by systematic counterfactual dependences between the neural realisers of M and B. If there are a great many neural realizers of M that enter into systematic counterfactual dependences with B, then it is very likely the realisers are values of a common neural variable that causes B. Let us call this neural variable N. I do not just mean that in the example $N = N_1 \vee N_2$. I mean that N is a genuine determinable that has N_1 and N_2 as determinates.

The argument for this conclusion gets even stronger when we consider not just the causal relationship between M and B but all the causal relationships M has with other variables. A generalisation of Figure 3 would show that the neural realizers of M must enter into systematic counterfactual dependences with these other variables in a way that mirrors their counterfactual dependences with M. That makes it even more likely, it seems to me, that these neural realizers are values of a common neural variable, N, that has the same causal role as M. But if N has the same causal role as M, then it is reasonable to think that N is contingently identical to M.

Does this mean that I must accept the principles *Top-Bottom* and *Bottom-Top*? The answer is 'No'. Even given the strong result just that M = N, N_1 is a determinate of the determinable N (=M), and we have seen that the contrastive account of causation allows that determinates can have causal roles independent of the causal roles of their determinables.