Can Structural Equations Explain How Mechanisms Explain?

Nancy Cartwright

UCSD and Durham University

1. What's in this paper

Peter Menzies has made a great number of important contributions to studies of causation over his career. Not only his specific ideas but his imagination, his approach, his excitement about the work, and his engagement with others and the way they think have had a huge influence in our field and especially on my own work ever since we first thought together about causality when Peter was a graduate student at Stanford. Here I shall discuss one of these contributions, a new one that Peter has developed very recently.

In 'The Causal Structure of Mechanisms' (Menzies 2012), Menzies deploys an interventionist account of causation to tackle the question of what a mechanism is, with special attention to Carl Craver's theory of mechanisms. Menzies has done so in his usual generous way, not by fitting Craver's mechanisms into a Menzies version of an intervention or manipulation account, but by following up Craver's own suggestion to use James Woodward's invariance-under-intervention account. In his own words, Menzies aims 'to show how the interventionist approach to causation, especially within a structural equations framework, provides a simple and elegant account of the causal structure of mechanisms.' (2012, 796) In particular Menzies wants to show *how* mechanisms explain the causal regularities they are supposed to explain.

I have for a very long time[1] been an advocate of just the kind of mechanism that Craver advocates, along with his sometimes co-authors Peter Machamer and Lindley Darden, as well as my colleague Bill Bechtel, and that Menzies is concerned with: 'a set of entities and activities that are spatially, temporally and causally organized in such a way that they exhibit the phenomenon to be explained',[2] where 'the aim of

[1] Cf. 'Where do laws of nature come from' (Cartwright 1997), reprinted in (Cartwright 1999). Except I have emphasized not the activities of the components, as Machamer, Craver, and Darden do, but rather their causal capacities

[2] Language here can be confusing since there are a handful of expressions in use in the causation literature that get assigned very different meanings by different researchers. I have used the word "capacity" to mark out something like J.S. Mill's tendencies: a feature has a stable capacity when it makes the "same contribution" to a specific kind of effect across a wide range of arrangements, regardless of what the actual "overall" effect is in those arrangements. I take this to be a central notion in any science that works by the analytic method, which assumes for many kinds of factors that what happens when factors operate in conjunction can be (at least in part) analyzed into the separate contributions that each make, which are just what each would produce were it to operate just "on its own". As we see in Section 2, Menzies too uses the word "capacity" but his expression picks out what I have always called (local) "causal laws" or "causal principles".

mechanistic explanation is […] to reveal the mechanism underlying [the phenomenon].' (Menzies 2012, 796) But I make almost the exact opposite claim to that of Menzies. I distinguish between the underlying structure, the mechanism -- which I have called a 'nomological machine', and the 'surface' phenomena that result when the machine operates. The interventionist approach to causation, especially within a structural equations framework, is not, I shall argue, at all well suited to represent the mechanism; but it can be very well suited to represent the causal regularities that the repeated operation of the machine gives rise to.

Menzies concentrates on Carl Craver's 2007 book *Explaining the Brain*, since there Craver provides such a fully developed account of the causal structure of mechanisms, and it is the details of the causal structure that Menzies is concerned with. Yet, Menzies thinks, Craver's account suffers from important omissions. It leaves the central notion of activity unelucidated and does not adequately show how the component entities and their activities are 'organized so as to exhibit the explanandum phenomenon'. Structural equations that satisfy interventionist criteria can do both jobs in one fell swoop, Menzies argues. I will sketch how he proposes to use interventionist structural equations to do so. Then I will explain why I have had a different view from the one Menzies defends. I shall argue that there is something absolutely essential that is still left out -- the very facts about the components and their organization that are responsible for the machine's capacity to produce the causal regularities we are trying to explain, which can often be well represented in interventionist structural equations.

In the end, though, I will admit that there is no fact of the matter. We can refuse my two-tiered account, with its distinction between, on the one hand, an underlying mechanism and its organization, and on the other hand, the surface phenomena, described in structural equations, that the mechanism gives rise to. We can, as Menzies advocates, use structural equations to represent the missing organization, and this can have some advantages, both for inference about the causal relations generated by the mechanism, and for purposes of systematic representation. But this representation is not transparent. If we do use it, the equations themselves do little of the work of representing the structure of the underlying mechanism; rather the important information gets buried in the description of the quantities that the variables in the equations are supposed to stand for. This strategy can also be dangerous in practice. To estimate these equations, or to confirm them, we have to measure the quantities represented by the variables. It becomes all too easy then to focus on the measurement procedures and to act as if these procedures identify the quantities. But they don't at all. Those same procedures will measure very different variables when applied in systems with different underlying mechanisms. We don't have to conceive of the situation in terms of mechanisms and their organization; we can stick with variables and equations. But without knowledge of the facts about organization and what it does that I take to be central to characterizing a mechanism, we don't know what variables we are talking about.

I shall also in the course of discussion endorse G.E.M. Anscombe's view that "cause" is a highly general notion, what I call a "Ballung" concept. I shall point out that these kinds of notions need to be made more precise if they are to play a proper role in scientific investigation and discourse and I shall defend Menzies's assumption that the intervention account provides sufficient conditions for picking out causal relations when what counts as a causal relation is precisely specified via a structural equations framework. This will, however, lead me to a minor disagreement with Menzies over the admission of transitivity as one of the characterizing features of a structural equations framework.

---

2. What is an activity?

As in his previous work with Peter Machamer and Lindley Darden (2000), in *Explaining the Brain* Craver takes activities to be central to characterizing mechanisms, where, as Menzies quotes, Craver uses 'the term "activity" as a filler-term for productive behaviours (such as opening), causal interactions (such as attracting) omissions (such as occur in cases of inhibition), preventions (such as blocking), and so on.' (Menzies 2012, 798) Menzies points out that in the 2007 book Craver adopts James Woodward's interventionist account of causation as an aid to explaining causal relevance, whereas Machamer, Darden and Craver (2000) 'endorse Anscombe's remark that the word "cause" is highly general and only becomes meaningful when filled out by other more specific causal verbs, e.g., scrape, push, carry, eat, burn.' (Menzies 20012, 798) Menzies praises Craver for going beyond these 'platitudinous remarks' by adopting an interventionist account.

Menzies's focus will be on causal relations expressed through functionally correct equations, where the effect is to be represented on the left and only causes of that effect appear on the right. At one stage Menzies uses these equations to discuss a case of singular causation but for the most part the discussion concerns causal regularities, following Machamer, Darden and Craver's claim (which he quotes) that the activities of a mechanism are 'productive of *regular* changes.' (Menzies 2012, 798, emphasis added) I call equations like this "causal principles" or (local) "causal laws". Menzies (2012, 800) calls them 'causal capacities' because, he explains, the kinds of equations in question imply 'a battery of interventionist counterfactuals', i.e. counterfactuals about what would happen if interventions were to occur on right-hand-side variables,[3] where, according to Menzies, 'Roughly speaking, an intervention on a variable X with respect to Y is a hypothetical experimental manipulation of X that is ideal for determining its causal influence on Y.' (2012, 798). The interventionist account then that Menzies subscribes to in this paper is like that of Judea Pearl and Woodward in demanding invariance under interventions on all right-hand-side variables: '… in order for these equations to capture causal relations correctly they must hold invariantly under interventions, or, in other words, the equations must continue to hold not only when variables on the right-hand side take on values in the normal course of events,[4] but also when these variables have their values set by a range of possible interventions.' (2012, 800)

An intervention is akin to the miracles that David Lewis uses to determine the truth values of causal counterfactuals -- the value of X is changed and only the value of X and the effects this change produces, leaving unchanged the remainder of other causes of Y (except for the downstream effects of changing X) as well as the causal principles at work (excluding the principles that govern the production of X). The main difference from Lewis is that for Woodward[5] somehow the change in X must be producible at least

---

[3] Interventions, being happenings in the world, affect quantities, not the variables that represent the quantities. But this longer expression is often cumbersome so I will, where little confusion could result, often just speak of variables for short rather than the quantities represented by variables.

[4] This mirrors my remark above that the equations are meant always to be functionally correct.

[5] Unlike for Pearl or Spirtes, Glymour, and Scheines, who can do with a purely conceptual notion of intervening: in an intervention a new system of causal equations replaces the old, but there is no need

in principle by some at least *possible* happening. Woodward calls his account indifferently both an intervention account and a manipulation account. It is presumably this insistence that there be at a least some possible happening that can change a cause in the right way that earns[6] it the 'manipulation' title[7]. If there is no possible way to change a factor on its own, then that factor can never be labelled a cause on Woodward's account, even should it pass other usual tests for causality (for instance, process tracing tests). Perhaps this aspect of Woodward's view makes it especially attractive for Menzies, who has long stressed the importance of manipulation to the concept of causation, though Woodward very much stresses, oppositely to Menzies, that manipulations in the required sense need have nothing to do with anything we conceive of doing.

Like Machamer, Darden and Craver I too have long followed Anscombe's view that the ordinary concept of "cause" is highly general. It is what, following Otto Neurath, I call a "Ballung" concept. A Ballung concept is a concept with rough, shifting, porous boundaries, a congestion of different ideas and implications that can in various combinations be brought into focus for different purposes and in different contexts. Many of our ordinary concepts of everyday life are just like this. Ballung concepts also can, and often do, play a central role in science and especially in social science. But they cannot do so in their original form. To function properly in a scientific context they need to be made more precise. This will be done in different ways in different scientific sub-disciplines, serving different ends and to fit with the different concepts, methods, assumptions and standards operating in these disciplines. The more precise scientific concepts that result will in general then be very different from each other and different yet again from the original Ballung concept.

I sometimes use the ugly word "precisification" to describe the process by which a Ballung concept is transformed into one fit for science. Sophia Efstathiou (2009) calls this process "found science" on the analogy of found art. Damien Hirst's shark in formaldehyde is still a shark but it is not the same shark as when it was swimming in the sea. It has been made suitable for an artistic context, to serve specific artistic purposes. In Efstathiou's words, the found shark has been "founded" -- given a form appropriate to serve its new purposes -- in the artistic context. But the shark now "founded" as art has lost many of its original functionings, including its ability to be founded in other contexts, such as shark soup.

So long as *causation* remains a Ballung concept, it is ill suited to serve scientific purposes. But it can be founded in various ways to make it more suitable -- causal pluralism in the flesh. As with the shark -- and as Efstathiou argues for other scientific concepts ("race" being one leading example), once causation is founded in one of these ways it can no longer carry out all of its original functionings. Importantly, the

---

for it to be possible to create a system satisfying these equations. They can be seen as a mere calculational device to fix the truth value of counterfactuals and new probability claims.

[6] It is unclear whether Woodward means 'there is something that is a possible happening that changes X in the right way' or 'possibly there is some happening that changes X in the right way'; nor is it clear that he would adopt the Barcan formula to move from the latter to the former. For more on the troubles with Woodward's notion of a possible intervention, see Marcellesi (Ms.).

[7] It is also this insistence that makes for the chief difference between Woodward's and the already long available accounts by Spirtes, Glymour and Scheines (2001) or by Pearl (2000) (if he is prepared to accept the causal Markov condition, as it seems he wishes to do) or by me (Cartwright 1999, 2007) (that does not suppose the causal Markov condition).

functions it can perform given one founding will not generally be available given other ways of founding it. That's where causal pluralism bites. The different foundings are not different ways of measuring or characterizing what remains the same concept. So what can be shown true of causation under one founding cannot be presumed true under another and empirical methods that work for telling where one obtains do not normally secure a causal relation that has been founded in any other way. In particular they do not license inferences that follow given other foundings.

Consider for example the set of relations that provide me with my morning toast,[8] relations that, as I argue in (Cartwright 2007), have familiar thick descriptions of the kind that Anscombe refers to: pressing the lever on my toaster *lowers* the spring-loaded rack where my bread sits, lowering the rack *closes* the circuit, closing the circuit *switches* on the heating element, the temperature rise *expands* a metal strip, … the movement of the catch *trips* a lever, the lever *releases* the toast rack, the rack *springs back*, *loaded* with the bread that has been *browned* by the same heating element that expanded the metal strip.

These are all good examples that fairly clearly fall under the everyday Ballung concept of causality. But that is not a sufficiently precise concept for use in science, especially not for precise prediction. For that we must found the concept more precisely and more explicitly. Perhaps, as I describe in (Cartwright 2007), we can classify the set of relations involved as causal in the sense prescribed by the axioms for causal Bayes nets, or in the causal structural equations sense that I shall describe, or what in (Cartwright 2007) is called "Hoover causality" after the economist and philosopher Kevin Hoover. But we must be careful not to pun. Generally a set of relations that is causal under one characterization will not be causal under another. In (Cartwright 2007) I provide one simple illustration where "mechanical causation" and "Hoover causation" give opposite verdicts about one and the same pair of event-types in one and the same machine. One has it that A causes B, the other that B causes A.

Of course this problem is not restricted to concepts of causality but is endemic throughout the sciences. Nor am I alone in my concerns. William Wimsatt, for example, makes a similar warning as mine against punning in science: 'Application of a heuristic to a problem yields a *transformation* of the problem into a non-equivalent but intuitively related problem. Answers to the transformed problem may not be answers to the original problem.' (2007, 346, emphasis original)

To return to mechanisms. Menzies takes the structural equations framework, which I shall explain in Section 3, to be the appropriate one for making precise what the causal principles (or in his terminology, the 'causal capacities') are that mechanisms give rise to.[9] He does so in part because it provides an answer to the question Anscombe, Machamer, Darden, Craver and I all leave unanswered: 'what [do] all these activities [like scrape, push, carry, eat, burn] have in common that makes them causal'? (Menzies 2012, 798) My own view is that there is nothing, and that that is not a problem. That's just what many of our everyday concepts are like: they involve a loosely connected set of ideas, different ones of which can be highlighted on different occasions to play different roles, from assigning moral or legal

[8] Cf. (Macaulay 1988, 159).

[9] Menzies calls these 'causal capacities', using the term differently from the usage I have made of that term in discussing mechanisms. (See note 2.) These are at any rate whatever it is that the causal equations represent.

responsibility to describing reasons for actions to providing advice about how to repair a system or how to avoid a catastrophe. Often they work like J.L. Austin's "trouser words": they get their sense in a context from what they are meant to rule out in that context.[10] So I think the attempt by philosophers to find something in common across all these activities is a mistake. In making this attempt we philosophers are 'bringing more rigour to a subject than it can bear'[11] (as we too often do).

In particular, the intervention account is stronger and more restrictive than other claims about the centrality of manipulation to the concept of causation that Menzies has developed and defended. Consider for instance his joint paper with Huw Price arguing that causation is a secondary quality like colour (Menzies and Price 1993). What makes a scraping a scraping or a pushing a pushing depends just on how the world is. What makes us label them both causings depends on our engagements with the world, though not, as with colour, via a particular sensory modality, but rather, in the case of causation, via our role as agents and as rational deliberators about how to achieve our goals. We strategize about the world, engaging in means-ends reasoning; the relations we label "causal" are ones we imagine could be used as strategies, or relations that we consider analogous to them with respect to the intrinsic properties that allow them to serve as means to the specified ends. Objects can sometimes be moved by pushing; surfaces cleaned by scraping. Other pushings get called causings by analogy. As Menzies and Price explain, on analogy with why we can claim that various objects that never could be viewed are nevertheless coloured, 'a pair of events are causally related just in case the situation involving them possesses intrinsic features that *either* support a means-ends relation between the events as is, *or* are identical with (or closely similar to) those of another situation involving an analogous pair of means-end related events.' (1993, 197, emphasis original) So Menzies and Price do not require that there be a possible manipulation in every case.

In addition it should be noted that the demand that a causal relation that a mechanism affords between X and Y should 'support a mean-ends relation' does not require that X appear in causal equations for Y deemed correct under the invariance-under-intervention account. Causal regularities can fail where there are genuine means-ends relations for a variety of reasons. Consider: X might be a means to increase Y but Y doesn't change because X is also a means to decrease Y in some other way and the two cancel each other. Then X will not appear (nontrivially) in many equations for Y deemed causally correct under the invariance-under-intervention account that Menzies and Craver adopt. Still we do in this case have two means-ends relations between X and Y -- and very often we need to know that. For instance, if Y is undesirable and there is a danger that the strength of the "decreasing Y" wing will diminish, knowing that X is a means to increase Y can help us plan to avert the bad effect of X on Y, or at least to prepare for it.

Another reason is that requisite supporting factors may be regularly missing so that a means-ends relation that the mechanism affords may never be realized, so no related causal regularity of the kind mechanisms are meant to explain actually obtains. For instance, I have bought a special flashlight for camping that will produce light by cranking a handle if, but only if, the battery is low. But also, I always check the batteries before setting out, so the handle never gets cranked. Still the fact that cranking is a

[10] Cf. (Austin 1962).

[11] As my late husband Stuart Hampshire put it.

means to produce light when low battery is added is a critical causal fact to know about this flashlight.[12] So when we are talking about criteria for the causal relations that a mechanism affords, very much depends on what kinds of causal facts we have in view.

The invariance-under-intervention account is also stronger than another view that Menzies is famous for: that "causes" is a theoretical term that, when correctly used of a pair of events in a situation, picks out a relation in the world intrinsic to those events in that situation but not one definable in non-theoretical terms.[13] The relation is rather picked out by a handful of platitudes that are often but not always true of it. One could even then take something like invariance under intervention to express one of these platitudes in cases where the relation could appropriately be represented in an equation of the type to which the invariance-under-intervention account applies.

As an aside to my main point here, I should like to note that this view of Menzies has been part of the inspiration for my own claims that "causes" is a Ballung concept; that when used in rigorous contexts, it needs to be specified precisely in a way that inevitably loses some of the features in the ordinary bundle of platitudes associated with it; and that when correctly applied, it refers to real causal relations in the world. I have not however tried to saddle Menzies with this inspiration since we have one big difference here. Menzies usually talks as if there is *one* intrinsic relation that is picked out whenever the term is applied correctly, whereas I (following Anscombe) think that there are countless different relations that we pick out in this way, which we also refer to with other more concrete, "thicker" descriptions like scrape, burn, push and eat.

Returning to the central point: The invariance-under-intervention account is stronger than either of these two accounts that Menzies himself has defended. In taking it up in his discussion of mechanisms, however, Menzies is following Craver's lead, since Craver himself suggest that the productive activities of mechanisms may be characterized using this intervention account. Perhaps then Menzies is not favouring intervention over his own previous views but exploring and developing Craver's idea, by adding on the structural equations framework to show how successful it can be in representing the causal structure of mechanisms. I shall proceed in the same manner. Structural equations may not be appropriate to represent all the different kinds of causal relations afforded by mechanisms, and they may not appropriate to all mechanisms, but they certainly are appropriate in a great many cases to represent the causal principles that describe causal regularities that could result from the repeated unimpeded operation of mechanisms of the right sort and that the mechanisms are supposed to explain.[14] In that case I also agree with what Menzies takes for granted, that an invariance-under-

---

[12] In my terminology, the flashlight has the stable capacity to produce light by cranking even if repeated operation of it never gives rise to a causal regularity exhibiting this capacity.

[13] Cf. (Menzies 1996)

[14] What sort is that? The answer is: what I call nomological machines. And what is a nomological machine? I answer to that in what I take to be a reasonable but unenlightening way: a nomological machine is an arrangement of features that have the capacity when operating repeatedly together unimpeded in that arrangement to generate the kind of causal regularity we record in a (local) causal principle, where I put "local" only in parentheses because I hazard that all, or almost all, causal

intervention criterion is appropriate to a causal structural equations framework, at least under some probably widely assumed assumptions about such a framework. That is what I shall explain next.

3. What is a structural equations framework?

First, we shall deal with linear equations only. I am fairly certain that the result I shall mention holds for non-linear equations, but I have not seen nor produced a proof. Linearity is not however such a strong constraint since products can be given a linear form by taking logs and also variables can be introduced that represent clusters of non-linear terms, though of course with great loss of information. I shall also deal only with deterministic equations since the understanding of what exactly the invariance-under-intervention account says about probabilistic equations is unclear.[15]

Menzies does not describe what a causal structural equations system is but it is clear both from his work and the account of Woodward that he endorses that it is meant to looks like this:

x_1 c= u

x_2 c= $a_{21}x_1$

x_3 c= $a_{31}x_1 + a_{32}x_2$

…

x_n c= $\sum a_{ni}\, x_i$.

The idea is that the equations in the system are supposed to represent true causal principles that hold for a given kind of situation, with effects on the left-hand side and causes, and only causes, on the right-hand side. The symbol "c=" expresses this asymmetry. There are a number of necessary conditions commonly assumed for equations of this form (linear, deterministic) to represent generic causal truths. The relations between right-hand-side variables (meant to represent causes) and left-hand-side variables (meant to represent their effects) are irreflexive and asymmetric. Third, any equation that is causally correct in a given kind of situation must be functionally true in that situation type. Fourth, the causally correct equations for a situation are the foundation for all other functionally correct equations that hold in that situation, in the sense that equations that are functionally true but not causally correct are obtainable by linear transformations and substitutions from those that are. This latter is the kind of assumption invoked when we suppose that spurious correlations, as between the fall of the barometer and the storm, must be explained by a common cause, like low pressure (or in some other way by

---

principles are local in this way to the operation of an underlying mechanism with the capacity to generate the regularity referred to in that principle.

[15] For one account of what reasonably might be intended see (Cartwright 2007, chapter 10).

reference to genuine causal principles). So, let x_1 = low pressure, x_2 = barometer drop, and x_3 = storm, and suppose the following causal structural equations system:

LP

1) x_2 c= $a_{21}x_1$

2) x_3 c= $a_{31}x_1$.

Then it follows that,

3) $x_3 = (a_{31}/a_{21})x_2$

-- a 'spurious' relation between joint effects of the common cause, low pressure, that is readily derivable from the two properly causal equations. These four are standard everywhere. They are either explicitly stated or implicit in the use to which these systems are put. Though not providing a reductive account of 'correct causal principle' together they constrain the notion, just as the intervention account of generic causal claims does. I add in addition an assumption that is widespread in use though is subject to controversy among philosophers, a requirement that I have called 'transitivity' of causal principles: an equation that results from substituting the right-hand-side causes of a variable in a causally correct equation for the variable when it appears as a cause in a causally correct equation is itself causally correct. I will discuss this further in Section 5.I have fleshed out the account of what a causal structural equations system is a bit more than Menzies but nothing I have said is incompatible with anything he says or uses in his discussion, with the exception of transitivity.

Menzies also, following Woodward, adds what they both call a 'modularity' assumption as a world-involving requirement on a correct causal equations set. Menzies explains modularity this way: '… a set of equations is modular if and only if it is possible to intervene on a variable on the left-hand side of an equation without disturbing the other equations in the set, i.e. without rendering the other equations false.' (2012, 800) For a fair test of a causal principle by intervention on its putative causes, it is clear that when those causes change value, other causal principles in the system should not be allowed to vary; otherwise we could be severing causal connections that a cause depended on to produce its change, or putting in connections that weren't there before that then bring about changes, or altering the functional form of the causal dependence. The modularity assumption does that job. But it is stronger than needed for just that job since it not only requires that what gets called an intervention leaves the other equations in the system intact. It also requires that such interventions are always possible on any quantity that get can get labelled a cause. This is just the requirement I have already made note of in Section 2. I myself argue that it is far too strong.[16] But that is not relevant to the issues about whether the causal structure of mechanisms should get represented as two-tiered, as I have been urging, or on one plane, as Menzies depicts, so I won't discuss it further. In what follows I shall not take modularity as a necessary condition on a correct casual equation system.

[16] Cf. (Cartwright 2007, ch. 7).

Suppose we adopt, as is common, the four necessary conditions I have laid out for a generic causal claim expressed in a linear deterministic equation to be correct, with or without my fifth sufficient condition of transitivity. This provides an informative characterization that constrains the undefined concept of a correct generic causal claim of this form. We also, though, have the invariance-under-intervention account that constrains the concept of a correct generic causal claim. As Menzies reports, 'Pearl and Woodward espouse the view that in order for these equations to capture causal relations correctly they must hold invariantly under interventions…' (2012, 800).[17] But exactly what is the connection between these two different characterizations? For instance, is invariance under intervention an additional requirement? If so, can it always be added consistently to the causal structural equations constraints?

Woodward defends application of his invariance-under-intervention account of causation to structural equations by example. For instance, if an intervention of the right sort occurs to x_2, which appears on the right-hand-side of equation LP 3), equation LP 3) will no longer hold: breaking the barometer will not bring on the storm, whereas an intervention on x_1 will leave both equations LP 1) and LP 2) invariant. Menzies provides similar examples to illustrate the invariance-under-intervention requirement. But it is not obvious that the two different criteria will always yield the same result for equations of the right form. Examples can illustrate how this works, but not show it. We can do better. It is possible to prove that if all a set of equations satisfy the four necessary conditions for being causally correct (and 'transitivity' holds for causal correctness), they also satisfy the invariance-under-intervention requirement.[18] This shows that the way in which Menzies marries the two in his discussion of mechanisms is entirely justified. Not only are the two sets of constraints consistent – the usual constraints for causally correct structural equations secure principle invariance under intervention.

4. How Menzies uses structural equations to describe mechanisms

What then does Menzies do with this apparatus? He uses it to explain what a causal mechanism is and what are and are not its constituents, and in such a way that the notion of activities is made sense of -- via the invariance-under-intervention requirement that ensures that the equations represent *causal* relations, not mere associations. Menzies supposes that the aim is to use a mechanism to explain what I

[17] Though, as Alex Marcellesi points out (in correspondence, 16 April 2013), it is not clear that Pearl insists that invariance implies causal correctness since he takes the idea of causally correct equations to be given, as the starting point for his analysis, whereas Woodward takes invariance-under-intervention to be the central informative characterizing feature of causal correctness. But Pearl certainly claims that correctness implies invariance and where invariance seems to be missing, the system must be misspecified.

[18] Cf. (Cartwright 2007, chapter 10). The reverse is also true if intervention is defined as I think it must be. My proof of these results assumes transitivity because, as I argue in Section 5, I suppose that is the right thing to do. Those who do not wish to assume this yet wish to marry a structural equation framework satisfying the four necessary conditions in general use to some other constraints on generic causal relations will need to produce their own proof that the two at least are consistent.

would call a generic input-output causal relation,[19] a relation describing a causal regularity generated by the repeated operation of the mechanism, of form O c= $f(I_1,\ldots,I_n)$. Menzies calls each structural equation a 'capacity'. As he explains, 'This terminology is motivated by the fact that a structural equation that is invariant under interventions implies a battery of interventionist counterfactuals.' (2012, 800) Then, '…any variable that lies on a pathway between the input variable and output variable of the capacity [generic causal relation] to be explained counts as part of the mechanism underlying the capacity [generic causal relation to be explained].' (2012, 801)

So what Menzies calls the *causal structure of the mechanism* that explains the input-output capacity [generic input-output causal relation] to be explained is a set of modular equations that first, constitute a causal structural equations system (so, each passes the invariance-under-intervention test) and second, compose to yield the input-output generic causal relation/capacity to be explained, where composition consists in substitution of the initial causes of later effects everywhere a later effect appears.[20] (Note: this is what I have called 'transitivity.) So, 'the *causal structure of a mechanism* is given by a set of modular subcapacities [generic causal relations] whose sequential exercise has the input-output profile of the capacity [generic causal relation] to be explained.' (2012, 800-801, emphasis original) This last is what solves the problem that Menzies worries about, of how the mechanism is supposed to explain the targeted input-output relation. Craver says that it is not covering law explanation. If not that, then what? Menzies answer is that the mechanism is a sequence of causal regularities that are instanced one after another, resulting in the regular causal connection between input and output. That, I take it, is the intended material mode version. In the formal mode, the equations representing this sequence of regularities compose via a sequence of substitutions to yield the equation that represents the causal regularity between inputs and outputs.

I want to suggest that this system of causal equations that describe activities or capacities that will be exercised sequentially between input and output does not represent mechanisms of the kind Craver and others and I have been concerned with (unless excessive work is done by the characterization of the variables in the equations). It does though represent a mechanism in the sense of the term often used in the medical literature: the step-by-step causal pathway that leads from the input to the output. That is the topic of Section 6. But first a diversion on transitivity.

5. Transitivity?

I suppose that causal correctness is transitive in the sense that taking the causes of a factor from a causally correct generic equation and substituting them for that factor where it appears in another causally correct equation yields an equation that is also causally correct. Woodward has used this fact to fault the proof that grounds the invariance-under-intervention account in an account that supposes the five conditions for causally correct generic equations that I have described here. He does so on the

[19] By calling it this I mean to imply that relations of this sort satisfy the five conditions for causally correct generic causal relations described in Section 3. I take it from all he says that Menzies' 'capacities' do so as well.

[20] So the principle to be explained is what econometricians often call the 'reduced form' of the system. Note too that what Menzies calls composition here is just what I call 'transitivity' in Section 5.

grounds that causation is not transitive. I want to discuss this briefly for two reasons. First, in defining mechanisms, Menzies also explicitly rejects transitivity. He says, 'A sequence of causal capacities [generic causal relations] described by structural equations do not by themselves constitute a mechanism.' (2012, 801). Second, the rejection of transitivity would undercut my defence of the way Menzies marries the invariance-under-intervention account with the causal structural equations framework (though perhaps another proof can be provided that does not suppose transitivity).

In reply I should like to point out first, that I don't see how an advocate of invariance-under-intervention as a characterizing feature of causation can deny this particular kind of transitivity since it is easy to prove that if the input principles pass the invariance test, so will an output principle that is derived by substitutions of the kind admitted by transitivity. I won't prove that here but will at least illustrate below with Menzies's own example. Second, the usual counterexamples -- and in particular those cited by both Woodward and Menzies -- involve singular causation, whereas the issue here concerns causal regularities. Whatever the case is with the former, exactly what is assumed in characterizing the latter is a matter of what more precise concepts find good uses in those settings where precise concepts matter.

Let me illustrate, using Menzies's example, the familiar case of 'boulder': enemy pushes boulder, walker ducks, walker survives. The pushing causes the ducking and ducking causes survival but pushing the boulder does not cause survival, or so intuitions seem to go. This is a case of singular causation, where each step is fixed. Yet Menzies, along with many others, writes it in terms of equations involving variables. His equations look like this:

BW

1) $P = 1$

2) $D = P$

3) $S = -P \vee D$

4) So $S = P \vee -P$.

Menzies maintains: 'But the enemies pushing the boulder does not causes the walker to survive. For whether or not the enemy were to push the boulder, the walker would survive. This is brought out by the fact that when we compose the structural equations, we obtain the result $S = -P \vee P$, which implies that S gets the value 1 whatever the value of P.' (2012, 801)

Menzies's version of the equations is a mix of Boolean notation and that of mathematical equations stating relations between variables. If we write the same information in pure equation form, we get:

BW'

1) $D \; c= \; P$

2) S c= (PxD) + (1-P) - (PxDx[1-P])

3) So S c= 1 - P + P^3.

4) P = 0 → S = 1

5) P = 1 → S = 1.

Notice I have not put in P = 1 because that is not an equation describing regimes of change among variables but rather the setting of a variable to a particular value -- presumably the value that variable actually takes on some occasion under consideration.

Equations BW'4) and 5) make clear Menzies's point though: S = 1 no matter whether P occurs or not. So equation BW'3 shows that the different values of what is pictured in it as a cause of survival do not make a difference to the value of survival. What's to notice though is that equation BW' 3) is invariant under interventions on right-hand-side variables. So if invariance-under-intervention is the mark of a correct causal principle, we had better let it in.

Of course we can stiffen the demands on our concept of 'causally correct generic relation'; we can add a requirement of change. I suppose it would be: If Q is a casually correct equation for effect y, then for each right-hand-side variable, x, in Q, there are at least two values X and X' for x and some arrangement of values for the other right-hand-side variables in Q such that y takes different values when x = X than when x = X'. Should one add such a requirement? Why? There is no right or wrong about the matter. These constitute two different foundings of the undefined concept 'causally correct generic relation' (or for Menzies, 'capacity'), one more restrictive than the other. There certainly is a use for the weaker way of founding the concept that does not demand that the effect value change with the cause value. Consider: if there are likely to be boulder pushers around, it is very important to know that the machine we have is designed to ensure the output *survival* regardless of whether a boulder is pushed or not. It is worth introducing the more restrictive notion if there are different sets of relations to which the stronger applies and sets to which only the weaker applies and we can do something with the information about which is which. What we must not do, however, is to ignore my warning in Section 2. We must not test for 'causes' in one sense, then draw inferences allowed only by a different sense without solid empirical evidence that the two co-occur in the kinds of cases where we draw the inferences. We must not do science by pun.

For purposes of thinking about Menzies' views on mechanisms, I think we can now set this issue aside. If a demand for change is not added, then invariance-under-intervention is a sure test that we have equations that satisfy the conditions sated for being causally correct, where transitivity is one of those conditions. If you don't like transitivity of generic causal relations, you will have to modify both the

assumptions I propose about them as well as add the requirement for change to the invariance-under-intervention account, and in a way that ensures the two line up as Menzies desires.

6. Is there more to mechanisms than structural equations reveal?

My answer to this question is "Yes". I have proposed some more serious scientific examples elsewhere, in particular of socio-economic machines to make clear that the machines that generate the kinds of regularities we record in our scientific principles need not be made of material parts. But here I shall illustrate with a more light-hearted example that makes the point very apparent, an example I take from (Cartwright and Hardie 2012, 77). You can look there for a picture (permission to reprint here costs a lot of money!).

When I want to sharpen pencils I don't crank a handle nor close a circuit on a battery-operated sharpener. I fly a kite. I can do it that way because I have a very special pencil sharpener, designed by Rube Goldberg. We can represent the generic input-output causal relation from my Rube Goldberg machine in the form Menzies suggests: S = K, where S and K are two-valued variables with S = 1 for pencils being sharp and O for not sharp, K = 1 for kite flies and K = 0 for not flying. (I shall use two-valued variables throughout for simplicity.) Let me tell you about the *causal structure* of this mechanism in Menzies's sense: 'a set of modular subcapacities [generic causal relations] whose sequential exercise has the input-output profile of the capacity [generic causal relation] to be explained.' (2012, 801)

The flying kite pulls open a door (D = 1; closed door, D = 0). The open door allows hungry moths to escape from a cage (M = 1; moths contained in cage, M = 0). The moths eat a flannel shirt (F = 1; shirt does not disappear, F = 0). Reducing the weight of the shirt causes a shoe to step on a switch (SH = 1; shoe not on switch, SH = 0). .. and many more steps till eventually a cage with a woodpecker under it lifts (C = 1; cage unlifted, C = 0) and a woodpecker pecks the pencil (W = 1; woodpecker doesn't peck the pencil, W = 0) resulting in a sharpened pencil. Each of these is a causal regularity that is instanced each time the kite flies and that can get represented in the form of a causal structural equations system of just the kind Menzies recommends. Here then are the set of modular subcapacities (generic causal relations) that exercise sequentially to generate the capacity [generic causal relation] to be explained:

RB

1) D c= K

2) M c= D

3) F c= M

4) SH c= F

…

11) W = C

12) S c= W

So S c= K, as required.

Now we know the step-by-step sequential causal process that results in kite flyings sharpening pencils. But we do not know the structure of the underlying machine that gives rise to this sequence.

Suppose you want to build a machine that affords the causal regularity recorded in S = K. There are an indefinite number of designs you could produce. Suppose you had a more ambitious aim: to build a machine that not only generates the input-output regularity S = K but the entire causal structural equations system recorded in RB. There are still an indefinite number of designs you could use. The machine that Rube Goldberg in fact designed is like this: the kite string goes under a lower pulley then up over a higher pulley and is tied onto a door that slides up and down easily, on a fine net cage full of hungry moths; the entire environment is safe for moths; the flannel shirt is attached to a string that runs over a third pulley with a shoe tied on the other end of the string that just balances the flannel shirt before the moths get to it; the shoe hangs immediately above a switch….eventually … a cage is raised from over a hungry woodpecker allowing the woodpecker to reach over to the pencil and peck it sharp.

This is the information that is still missing, even once we have recorded the sequence of causal regularities that produce the overall input-output relation. It is the reason that I urge that we conceive of two tiers: the underlying arrangement of parts with their associated features and capacities -- ‘capacities’ in my sense; and the surface causal principles that are afforded by the underlying structure. Both are causal, in some sense, and both kinds of information are important to know. Both are necessary for a full understanding of how the pencils get sharpened. And both are useful for prediction and for means-ends reasoning.

First consider the surface equations. The other day Lucy was playing games with me and put her finger on the door to keep it closed every time I went out to fly the kite. She knew I would get no sharp pencils then. She knew that because she could read it off from setting M = 0 in RB 2) and following through the

downstream effects. One day later in the month the moths were gone from the cage because Lucy had taken them out the night before to see if they really were attracted to candle flames. We knew no sharp pencils that day either, from RB 3). Or on another day I couldn't get the kite to fly because there was no wind. And anyway I was in a hurry. So, guided by RB 11) and 12) I went straight to the cage over the woodpecker and lifted it. The point is that these all reflect proper capacities in Menzies's sense [generic causal relations in my terminology], passing the invariance-under-intervention test and thus supporting 'a battery of interventionist counterfactuals' (2012, 800) as Menzies wishes.

Now consider the underlying structure. One day I flew the kite but the door didn't open. I had been really cautious against Lucy's tricks and all other such hazards and had locked the machine up well the night before. So I was sure no external intervention had set any of the variables in RB to 0. I knew the machine must be broken and I would have to look inside to fix it. Indeed, the top pulley had split in half. I knew to check on the pulley because I understood the parts of the machine and how they worked together to afford the generic causal relations I usually could rely on to get sharp pencils. Or, more recently I decided I was bored by this machine and would like another, but one that sharpens faster and more evenly. So I have been busy reviewing how gears work and trying to figure out how to hook a knife blade to a windmill. The point here is that with knowledge of the parts that compose a machine and how they operate, we know what it takes to repair it. And with ingenuity and knowledge of the capacities [in my sense] of lots of different kinds of parts, we can build entirely new machines, hopefully better than the old.

7. Putting it all in one flat plane

You don't have to conceive of machines in terms of two-tiers, as I recommend. There are a couple of fixes if you do want to represent them on one flat plane.

First you can complicate the causal equations, like those in RB, by adding a new variable that takes value 1 when the machine parts and activities are all in place and working properly and 0 otherwise. Call this variable RB-NM (*RB* for Rube Goldberg, *NM* for my term -- 'nomological machine'). Then you multiply by RB-NM each cause in each principle in the original surface system of equations.

Some machines allow a more detailed breakdown: those that are modular, but in a different sense of 'modular' than that of Woodward and Menzies described in Section 3. A machine is modular in this alternative sense when input from separate causes depends on separate parts. Then each cause gets multiplied by a yes-no variable representing the proper functioning of the associated part. This is not an efficient design for a machine though since it means there are a lot of parts and the machine may then grow big and unwieldy. But it can have advantages. For instance, it will have advantages when we want to design the machine so that it will be easy to trouble-shoot. Perhaps the parts of the machine are difficult to access, so we want to make it easy to discover which piece is broken based on the surface behaviour of the machine; or when parts are likely to wear out and we do not want too many principles compromised at once, or when we think we may be able to secure one or another broken link with

some different part. Apparently this was one of the demands of the MIT WWII radar project because it was expected the radars would have to function in isolated environments and it was important to make repair as likely as possible.[21]

This fix does not of course do away with the need for exactly the same information required for the two-tiered picture. It just allows us to use a single representational scheme. This is fine so long as we do not suppose that all the variables in the equations have the same status and so can be treated in the same way. For instance, often because of the kinds of underlying machines that generate the surface equations, the surface variables satisfy the conditions for random variables, in particular, there is a probability measure over their values. But generally there isn't any probability for whether a machine of a specific design will be built or not, in which case RB-NM is not a random variable. So, too, often with the 'variables' that would appear in equations from machines with more modular structures.

A second fix is to add the information about the parts and activities of the machine into the variables themselves. So, for instance, in the system RB, K would no longer represent the feature 'Kite flies' but rather 'Kite attached to a string that goes under a low pulley and over a high pulley before attaching to the top of a little door flies'. And so forth. This too has problems, especially in practice. For one, as with the first fix, in this case too we can no longer assume that the variables in our new equations will be random variables, even if those in the original surface equations were.

Another problem has to do with how we connect our measurement procedures with the variables we measure. Whether a kite is flying is easy to measure. What about measuring whether or not a kite affixed to my Rube Goldberg pencil sharpener is flying? That is a different matter and requires hugely more information, information that we seldom have. Nor do we need it in order to confirm or estimate the surface equations -- and recall, we do want to know these surface equations because they show us effective means for achieving our ends. All we need to be sure enough of is that the data we gather is all generated by the same mechanism [in the sense of a nomological machine]; and there are a lot of things that can provide good assurance of this far short of having a complete description of the mechanism and checking through each of the details. That's a danger in one direction: that we may lose power to discover useful principles because we do not know what our procedures measure when these very complex variables are in play.

There is also a danger in another direction. Since we can perform the simple operations it takes to tell if a kite is flying and to tell if pencils are being sharpened no matter what mechanisms [nomological machines] are involved, we can lose sight of the importance of the mechanism. This I think happens regularly in evidence-based policy nowadays. There are now a great many agencies, like the US Department of Education's What Works Clearinghouse, that publicize 'What works', from educational to health to criminal justice to international development.

[21] Cf. (Galison 1997).

How do they decide what works? They look at scientific studies, where they are all keen to ensure that the studies are of the right kind and right standard to establish a genuine generic causal relation ['capacity' in Menzies' terminology]. So their basis is composed of studies, like controlled trials, well designed to test claims about generic causal relations/capacities -- describing regularities of the surface variety -- in one or two or a handful of sites. If a causal relation is established between two surface variables, like kite flying and pencil sharpening, or, more seriously say between hot spot policing and burglary reduction,[22] the cause gets recorded under the heading 'What works'. Then policy makers are advised to consult the relevant clearinghouse and to use only policies listed there under 'What works.' This seems to suggest that a cause known to have produced a desired outcome in a handful of settings will work in new places unless something special goes wrong, or that the assumption that it will work in a new place is the default assumption. But when, as is typical, the generic causal relations under consideration are surface relations, whether a proposed cause will work in a new place depends on whether the new location has the right underlying structure to support the same causal relations. But no-one says that finding a policy in a 'What works' list gives you negligible reason to use it if you have no information about the underlying mechanisms [nomological machines] needed to produce the causal relations you'd be relying on. The two-tier picture keeps this firmly in view.

8. The job done and the job to be done

One of Peter Menzies's aims in offering the structural equations system characterization of a mechanism was to show how the mechanism explains the input-output relations it is supposed to. When a mechanism is conceived, as Menzies urges, as the sequence of causal regularities instanced in between cause and effect, the goal has been achieved: the in-between regularities are expressed in structural equations and they explain the input/output regularity by implying it, via composition.

But there is still another layer of explanation, I have urged, and another sense of mechanism that does the explaining: mechanisms in the sense of the underlying arrangements that give rise to the entire set of regularities recorded in the causal structural equations system. We still need an account of *how* the underlying arrangement explains these principles. Nor is this problem an artefact of the two-tier picture; it simply appears in another guise if you look at everything on one flat plane, say by using variables that refer to the machine structure. In that case the causal relations we are considering will be very unfamiliar when written out fully; and there will be innumerably many of them, considering all the mechanisms that might occur in nature and society. Where did these all come from? They certainly won't look like the kinds of things we are used to thinking that God wrote in the Book of Nature. I do not think we have a good answer. That's why the question that Menzies has tackled is so pressing.

22 Cf. http://www.hmic.gov.uk/pcc/what-works-in-policing-to-reduce-crime/ (accessed 25 April 2013)

I have offered an answer for some special kinds of cases but my answer is not all that good. I start from my view that many of our basic so-called 'laws of nature' are best seen as ascriptions of capacities [in my powers-like sense] to features independently identifiable: like assigning to the feature of having mass M the capacity of strength GMm/r^2 to attract other masses. The capacity itself is identified not by what effect results when it activates but by the contribution it makes to the effects that actually occur, where it is supposed that the capacity makes the same contribution across a wide range of circumstances. In some nice cases a mechanism [nomological machine] will consist of parts with features that carry capacities for which there is a rule of composition about what happens when the capacities all contribute together. The well-known example is vector addition, which is how the contributions from various sources of attraction and repulsion combine. I have also described a number of other rules of composition we find in other disciplines for the capacities they study.[23]

In these cases the *how* question has an easy answer. A mechanism [nomological machine] explains the resulting causal regularities in that those regularities can be derived from the facts about capacities [in my sense] associated with the features of the mechanism plus a rule of composition. In paradigm cases the capacity claims employed as well as the rule of composition do look like what we have pictured God to write in the Book of Nature, since these will be more familiar 'laws of nature', like the law of gravitational attraction, Coulomb's law of electromagnetic attraction and repulsion, and vector addition; or the laws of simple machines and how the combine.

The trouble is that not many of the underlying mechanisms [nomological machines] we find on offer in the natural and social sciences to explain generic causal relations local to those mechanisms can be cast into this simple form. Even if my answer were good enough for these special cases, it just doesn't go very far. And Menzies is right to have underlined the question. I have not seen an answer yet that works. In offering the structural equations framework, Menzies has succeeded in answering the important how question for one sense of mechanistic explanation. But what are we to say about *how* familiar mechanisms from toasters to socio-economic structures explain generic causal relations, like pressing on the lever will brown the bread or installing CCTV camera will reduce car crime?

Acknowledgments

I would like to thank Alex Marcellesi and Gil Hertshten for research assistance as well as the Templeton project 'God's Order, Man's Order and the Order of Nature' and the AHRC project 'Choices of evidence: tacit philosophical assumptions in debates on evidence-based practice in children's welfare services' for support for research for this paper.

References

[23] See (Cartwright 1999, chapter 3).

Austin, John. (1962). *Sense and Sensibilia*. Oxford: Oxford University Press.

Cartwright, Nancy. (1997). 'Where Do Laws of Nature Come From?', *Dialectica*, 51(1): 65-78.

------. (1999). *The Dappled World*. Cambridge: Cambridge University Press.

------. (2007) *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Hardie, Jeremy. (2012). *Evidence-Based Policy: A Practical Guide to Doing it Better*. New York: Oxford University Press.

Craver, Carl. (2007). *Explaining the Brain*. New York: Oxford University Press.

Efstathiou, Sophia. (2009). *The Use of "Race" as a Variable in Biomedical Research*. PhD Thesis, UC San Diego.

Galison, Peter. (1997). *Image and Logic*. Chicago: University of Chicago Press.

Macaulay, David. (1988). *The Way Things Work*. Boston: Houghton Mifflin.

Machamer, Peter, Darden, Lindley, and Craver, Carl. (2000). 'Thinking about Mechanisms', *Philosophy of Science*, 67(1): 1-25.

Marcellesi, Alexandre. (Ms.). 'Interventions, counterfactuals, and causation: Some unfinished business', unpublished manuscript.

Menzies, Peter. (2012). 'The Causal Structure of Mechanisms', *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4): 796-805.

Menzies, Peter, and Price, Huw. (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science*, 44(2): 187-203.

Wimsatt, William. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge (MA): Harvard University Press.