Why do social scientists disagree about the right way to do causal inference?

Reading guide: *While the whole file is 36 pages long, pages 26 to 36 contain the bibliography and a technical appendix which can be skipped when reading. So, the final paper is 25 pages. If you find this sample too long, please feel free to skip Sections V.b-V.e. Thank you for your patience and I apologize for the inconvenience.*

Abstract

There are irreconcilable differences between two main approaches to causal inference in social science: Rubin's Causal Model (RCM) and Pearl's Causal Model (PCM). In this paper, I argue that these methodological differences can be largely explained by their different metaphysics of causation. Specifically, I show that the biggest methodological point of contention between RCM and PCM concerns whether we can estimate individual causal effects in actual causation causes. Building on this insight, I show that both RCM and PCM assume a reduction of causation to structural equations. However, on RCM these equations can *differ* across individuals whereas on PCM they are the *same*. I then suggest how this seemingly small metaphysical difference can explain most methodological differences between the two approaches.

I Introduction

There is a fierce debate about how to conduct causal inference in social science (Pearl, 2012; Imbens, 2020). On one side are the advocates of Pearl's Causal Model (PCM) who use graphs to represent the causal links between different variables.[1] On the other side are advocates of Rubin's Causal Model (RCM) who examine real-world cases which approximate controlled experiments. So, why do they disagree about causal inference? I answer this question in two steps. First, I show that there are irreconcilable methodological differences between PCM and RCM. Second, I argue that these differences can be attributed to the fact that they assume two slightly different metaphysics of causation.

[1] I largely focus on Pearl's exposition of PCM instead of alternative accounts, e.g., Spirtes et al. (2009) because I am more familiar with Pearl's work. In future work, I hope to examine if my arguments apply to other formulations of the graphic approach.

RCM advocates disagree with PCM advocates about many things and usually have a good response to any objections from their own standpoint, and *vice versa*. For example, when PCM advocates object that RCM ignores causal transmission mechanisms, RCM advocates can reply that this is actually a benefit of their approach. I will show that most such methodological disagreements reduce to a disagreement about the so-called *fundamental problem of causal inference*. Philosophically, this refers to the question whether we can estimate individual causal effects in cases of actual causation.[2] PCM advocates believe that this is possible which RCM advocates deny. Despite the importance of the *fundamental problem*, the debate cannot be resolved on purely methodological grounds.

Should we give up on resolving the disagreements? Luckily, the metaphysics of causation can tell us not only why such methodological differences exist but also why PCM and RCM work in practice. Drawing on Papineau (2021), I argue for reducing causation to a system of structural equations, i.e., the structural equations theory of causation. Papineau claims that his version of the theory is sufficient to explain why PCM works. I show that this is *not* always true, given the functional forms which he assumes on the structural equations underlying his theory.

For this reason, I modify his theory in two ways. First, I propose a restricted set of functional forms that are sufficient to explain why **P**CM works. Second, I suggest how to change the theory, so that it can also explain why **R**CM works. I then argue that RCM and PCM rely on two slightly different versions of this theory: on PCM the structural equations are the *same* for all people whereas on RCM they *vary* across individuals. This feature allows me to explain why RCM and PCM see the *fundamental problem* differently and, consequently, many other contentious methodological points between them.

To support this argument, I begin by describing PCM (Section II) and RCM (Section III). Next, I discuss the methodological debate between RCM and PCM (Section IV). I then show why the traditional metaphysical theories of causation cannot justify why RCM and PCM work (Section V). Last, I outline the structural equations theory of causation and suggest how it can explain the methodological differences (Section VI).

[2]This definition of the *fundamental problem* has a formal equivalent in RCM (Section III).

II Pearl's Causal Model

II.a Set-up

How does PCM-based causal inference work? Let us consider a simple example before discussing the more fundamental assumptions of PCM. Suppose that we are interested in the effects of neighborhoods *early* in life (N), i.e., the area in which a person was brought up, on income *later* in life (I).[3] Different authors have stressed different channels for neighborhood effects (Chyn and Katz, 2021). Based on such research, we can argue that neighborhood effects on income are fully mediated via education (E) and health (H), so that neighborhoods have no independent effect on income.[4] Figure 1 summarises these considerations. Given some additional assumptions discussed below, we can interpret the graph as causal.

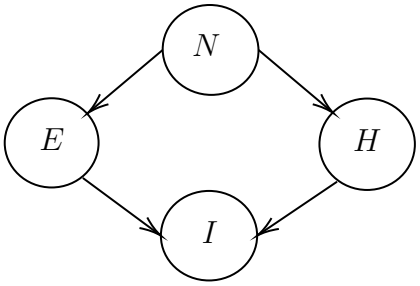

Figure 1: Neighbourhood effects on income

Before proceeding, it is useful to introduce some graph theory terminology. Graphs in this paper represent how a set of variables is connected. Each node represents a different variable, e.g., $\{N, E, H, I\}$ in Figure 1. Nodes are connected by *edges*. An edge is *directed* if it points from one node to another node. For any directed edge $N \to E$, we say that N is a *parent* of E and E is a *child* of N. More generally, all the predecessors of a node E, including its parents, are its *ancestors* whereas all its successors, including its children, are its *descendants*. Any collection of one or more consecutive edges is called a *path* where a *directed* path is a path containing edges directed only in one direction. A graph is *acyclic* if there is no directed path that starts from node N and ends in the same node, e.g., $N \to E \to I \to N$. These definitions

[3]Neighbourhoods are a very popular topic in social science (Sampson et al., 2002; Chyn and Katz, 2021). There is also substantial evidence for their importance (Chetty et al., 2016).

[4]Robert (1999) established $N \to H$, Kauppinen (2008) found $N \to E$, Card (1999) discussed $E \to I$ and Baicker and Finkelstein (2011) suggested that health insurance increases financial stability ($H \to I$). Other potential mediators, e.g., crime (Kling et al., 2005), are omitted for brevity.

are important since Pearl is largely interested in Directed Acyclic Graphs (DAGs) which is a necessary assumption in many of his theorems.

Without additional assumptions, DAGs do not necessarily have a causal interpretation: they could simply be summarising a probability distribution in a succinct way as in non-causal Bayesian nets. How can we then give Figure 1 a causal reading? While there are other options,[5] the approach favored by Pearl assumes the existence of a system of structural equations, in which the variables on the right-hand side directly determine the variable on the left-hand side:

$$y_k = f_k(pa_k, u_k)$$

where pa_k is a vector with the observed parents of y_k, k refers to the k-th equation in our system, u_k is an probabilistically independent error term capturing unobserved factors and $f_k(.)$ is some unspecified functional form. Crucially, while pa_k affect y_k deterministically, probability enters y_k via the errors u_k. In our example about neighborhoods, the complete system is:

$$Education = f_E(Neighborhood, u_E) \tag{1}$$

$$Health = f_H(Neighborhood, u_H) \tag{2}$$

$$Income = f_I(Education, Health, u_I) \tag{3}$$

where the u-s are probabilistically independent errors. Once we have our structural equations and give them a causal reading, they can be represented in a DAG.

II.b Causal inference with PCM

We have seen Pearl's view of the theory underlying the DAGs. How can we use the DAGs for causal inference? Here I focus on one aspect of causal inference: causal discovery, i.e., how to uncover the true causal structure from data.[6] To conduct causal discovery, we need to make

[5]The main alternative assumes probabilistic causation and provides conditions, under which DAGs can be read as *causal* Bayesian nets (Spohn, 2009). Pearl argues against this approach in Chapter 1 of his book (2009).

[6]Another use of PCM in causal inference is for *causal effects identification*. In the context of PCM, identification tells us if we can estimate a particular causal effect given our graph. I do not discuss this topic because identification methods are valid only if the underlying DAG captures genuine causal relations, i.e., discovery seems more important than identification.

certain assumptions about the true causal model. Studying such assumptions will allow us to make claims of the type: 'Suppose we have a set of correlations. If assumptions 1), 2), etc hold in the true causal model, then we can use our correlations to discover the true model'. If we do not make any assumptions on the true causal structure, we will be unable to determine whether a *hypothesized* model based on our evidence is approximating well the *true* structure. In other words, the assumptions allow us to say how good the model we have discovered is. The trouble is that many such assumptions are empirically unverifiable which makes causal inference difficult.

We can illustrate this idea by examining two such assumptions: *faithfulness* and *causal sufficiency*. First, *faithfulness* states that the absence of a causal link[7] implies the absence of a correlation (Zhang and Spirtes, 2016, p.1114). If we find no correlation between two variables (potentially after conditioning), they are not causally linked directly. There are various counterexamples to faithfulness, in which two variables have zero correlation but are actually causally linked (Andersen, 2013; Papineau, 2021). Here I provide a verbal version of such a counterexample and Appendix A.a provides the graphical equivalent.

Suppose we are interested in the gender wage gap and find that conditional on occupation there is zero correlation between gender and income. This is plausible given evidence that women tend to choose lower-paid occupations (Blau and Kahn, 2017). Does that mean that gender has no causal effect[8] on income outside of occupation? Not necessarily. Gender might still affect earnings through two separate paths which exactly offset each other, leading to a zero correlation. For example, note that women are more likely to attend university (Goldin et al., 2006) and that education increases earnings (Card, 1999). So, gender has a *positive* effect on income via education. At the same time, there might still be gender differences in wage negotiating ability (Card et al., 2016). As a result, gender may have a direct *negative* effect on income which *exactly* cancels out the *positive* effect mediated by education. This is a faithfulness failure: gender and income are conditionally uncorrelated, albeit gender still affects income. We can now see why faithfulness is problematic (aside from being unverifiable[9]). If it fails in the true model, any

[7]Two variables V and W are causally linked if V (in)directly causes W or if W (in)directly causes V or if they have a (in)direct common cause U (Papineau, 2021, p.8).

[8]For brevity, I assume that we can measure the effect of gender with PCM. I only make this assumption for simplicity but there is good reason to think it is false (Kohler-Hausmann, 2018; Hu, 2022).

[9]One might respond that faithfulness still has testable implications (Zhang and Spirtes, 2008). However, the fact remains that the assumption cannot be directly verified.

hypothesized causal links, which assume it, might not be truly causal.

The next assumption is *causal sufficiency* which allows us to rule out the effects of unobserved causes (Scheines, 1997; Papineau, 2021). If causal sufficiency holds, then we can never find a common cause of both variables in any hypothesized causal link from our model. The assumption does *not* require the inclusion of all variables affecting *the main variables of interest* but only variables affecting *both variables in a given hypothesized causal link*. In Figure 1, we do *not* have to include gender if it affects income directly but is not a common cause in any of the hypothesized causal links. We do need to include parental income if it affects both neighborhoods and income. If causal sufficiency is incorrectly assumed for a causal link, this link will not be truly causal. While there are discovery algorithms for datasets that do not satisfy causal sufficiency (Spirtes et al., 2009, Chapter 6), they are much less helpful, e.g., their output is not necessarily a DAG (Pearl, 2009, §2.6.). However, it might be that faithfulness and causal sufficiency actually hold in the true model.[10] In that case, there is a plethora of algorithms that are guaranteed to discover the true structure, e.g., Pearl (2009, p.50).

III Rubin's Causal Model

The basic primitive in RCM are the *potential outcomes* (POs) (Holland, 1986; Imbens and Rubin, 2015). Drawing on randomized control trials, POs are usually defined relative to a binary *treatment*. More formally, the POs for unit j are:

$$Y_{j,t}(D=1) \quad and \quad Y_{j,t}(D=0)$$

where D indicates if j got the treatment. In the example from Figure 1, the treatment could be living in a good relative to a bad neighborhood whereas $Y_{j,t}$ is j's income. We also assume that j's POs at time t are not the same as j's POs at time s (Imbens and Rubin, 2015, p.8). Intuitively, this is because the effect of an Aperol Spritz on my happiness in the *morning* is not the same as its effect in the *evening* after a long day of work.

[10] Technically, we also need the *Causal Markov Condition* and *minimality*. See Stern and Eva (2022) for details on these assumptions.

A crucial feature of POs in RCM is that treatment D can only be a variable, which we can imagine manipulating in a hypothetical experiment (Imbens and Rubin, 2015, p.21). There is 'no causation without manipulation' (Holland, 1986). This rules out studying treatments such as race because we cannot imagine designing an experiment in which we change somebody's race. While this is partly a matter of degree, it is much easier to imagine giving a person an aspirin than changing their race. RCM also allows interventions that influence race perception, so long as they are easy to implement (e.g. Bertrand and Mullainathan, 2004). Following Weinberger (2022), we should see this manipulability requirement as a constraint on the type of interventions that we can study in RCM, even though mathematically nothing prevents us from ignoring it.

If we decide to use POs directly to estimate individual causal effects, we will face the *fundamental problem of causal inference*. For each j, we observe either $Y_{j,t}(D = 1)$ or $Y_{j,t}(D = 0)$ but not both. So, we cannot calculate the *individual* treatment effects. Below, I argue that the *fundamental problem* is crucial for understanding the methodological differences with PCM. For example, it explains why RCM methods work under a lot of variation in individual effects (Imbens and Wooldridge, 2009, p.7) at the expense of ignoring mechanisms.

The *fundamental problem* translates to the philosophical literature on causation as the question if we can ever know the causal effects in a particular situation, i.e., actual causation cases.[11] For the remainder of this paper, I will refer interchangeably to the philosophical and the RCM formulation unless stated otherwise. Luckily, despite the *fundamental problem*, we can still get an unbiased estimate of the *average* treatment effect in randomized control trials, e.g., Theorem 6.1 in Imbens and Rubin (2015). This quantity is still useful as it tells us how people in our sample respond on average to the treatment.

Very often, however, we cannot conduct such experiments (for ethical or legal reasons) and need to rely on observational studies. It is, therefore, important to study the assumptions required for unbiasedness. This will allow us to check if they hold in a particular observational study and if we can use it to learn a causal effect. The stable unit treatment value assumption (SUTVA) is one such assumption (Imbens and Rubin, 2015, pp.9-13). One part of SUTVA requires that whether j is assigned treatment should have no bearing on whether i is assigned

[11] There are many other ways to understand actual causation (Russo and Williamson, 2011) and my view seems most similar to Papineau's single-case causation (1986, p.118 and p.123).

treatment, i.e., j's treatment should be independent of i's treatment.[12] While this assumption is not necessary for treatment effect estimation,[13] it is often imposed due to data limitations, similarly to faithfulness in PCM.

Other important assumptions concern the treatment assignment mechanism. Specifically, the assignment mechanism in randomized control trials in contrast to observational studies is controlled. This ensures that treatment is genuinely allocated at random, so that we can get an unbiased estimate of the average treatment effect. In observational studies without randomization, we can only get the causal effect if we make some additional assumptions on the assignment mechanism. One way to proceed is to assume selection (into treatment) on *un*observables (Cerulli, 2015).[14] On this assumption, treatment assignment depends not just on observable characteristics but also on unobservable ones. So, we cannot use standard techniques such as linear regression to estimate the causal effect, as we will run into omitted variable bias.

In response to this problem, methodologists have developed techniques such as instrumental variables, difference-in-difference, and synthetic control that work under selection on unobservables (Athey and Imbens, 2017). While such techniques require additional assumptions, they show that RCM causal inference can work even if we do not observe all variables that affect both treatment and outcome. This key feature also explains how Chetty and Hendren (2018) can estimate neighborhood effects when treatment is not randomly assigned.

IV Methodological disagreements

We have now introduced two seemingly different approaches toward causal inference. There is a lively debate between their advocates (Pearl, 2012; Gelman, 2009; Imbens, 2020; Heckman and Pinto, 2022). In this section, I contribute to a recent literature in philosophy of science that tries to uncover the main points of contention in this debate (Markus, 2021; Weinberger, 2022). In particular, I discuss two key clash points (mechanisms in RCM and causal sufficiency in PCM) and then consider how the two sides can be reconciled. Unfortunately, I show that this is not

[12] I do not discuss SUTVA's homogeneous-treatment part because of space constraints. See Imbens and Rubin (2015, pp.9-13) for details.

[13] A burgeoning literature is designing RCM method for situations when SUTVA fails (DiTraglia et al., 2020).

[14] The main alternative is selection on observables which requires that treatment assignment is fully determined by observable variables. Unfortunately, this assumption is known to be almost always false (LaLonde, 1986).

possible because of the *fundamental problem of causal inference* which can also explain why they disagree about many other methodological issues. In that sense, the *fundamental problem* is crucial for understanding the methodological debate.

IV.a Clash points

What do PCM and RCM advocates disagree about? One standard objection against RCM is that it tells us nothing about transmission mechanisms (Deaton and Cartwright, 2018). If we use RCM, we do not learn how a cause leads to a particular effect.[15] An RCM advocate might respond that this is actually a key strength of RCM. While sometimes there might be a clear mechanism behind a causal link, very often there might be too much variation across people. So, imposing a single model will be inappropriate or even wrong. This does not mean that we can never model the mechanism but only that modeling it correctly might be more difficult than assumed by PCM advocates. If we believe that methodology should be driven by pragmatic considerations (Weinberger, 2022), there is nothing wrong in using different methods to solve different problems. We can first use RCM to establish a causal link and then use other techniques to study the underlying mechanism. On its own, RCM does not exclude such an approach.[16]

On the other hand, RCM advocates can argue that PCM rarely identifies[17] causal effects, since causal sufficiency usually fails. Causal sufficiency requires that we include all variables that are common causes of the two variables involved in the causal link of interest (Section II). Sufficiency becomes problematic when *unobservable* variables affect our causal link because we cannot verify it empirically. Unfortunately, in social sciences 'any effects that can possibly be there typically are' there (Imbens, 2020, p.1140), meaning that it is usually easy to find a common cause of both variables in a causal link. We omitted criminal activity from Figure 1, although it might affect both neighborhood quality and income (Kling et al., 2005). So, causal

[15]PCM advocates might also argue that RCM lacks external validity. Setting aside the question of whether 'external validity' is a coherent notion (Deaton and Cartwright, 2018), it is not true that all PCM methods have more external validity than all RCM methods, e.g., most RCM-based research utilizes observational data. So, this critique applies to both approaches.

[16]One might wonder if PCM advocates will object to other assumptions of RCM. They cannot object to SUTVA since it is also an issue in PCM (Ogburn and VanderWeele, 2014). A more controversial assumption is consistency which seems to be differently understood in the two frameworks (Weinberger, 2022, §3).

[17]Identification is an elusive notion (Lewbel, 2019). In the context of RCM, I understand identification as the assumptions required to get an asymptotically consistent estimate of a causal effect from data via a particular causal inference method.

sufficiency probably fails and neighborhood effects are not identified.

PCM advocates may respond that causal sufficiency is *always* needed: RCM-based techniques will also fail unless we assume causal sufficiency. Nevertheless, causal sufficiency in RCM is less demanding than in PCM. The reason is that it can be translated into PCM as an assumption called selection on observables which requires that treatment assignment is fully determined by observables (Dale and Krueger, 2002). However, we saw that many RCM methods do not need this assumption: they are specifically designed to work under selection on *un*observables, i.e., when causal sufficiency fails. Furthermore, RCM usually explores the causal links between *two* variables only whereas PCM often explores the links between *many* variables. Why is the number of variables important? If we only study one causal link as in RCM, we should ensure causal sufficiency only for that link. If we study many causal links as in PCM, we should ensure causal sufficiency for all links. So, in PCM we need to guarantee causal sufficiency for many more links than in RCM. Unfortunately, this is difficult even for one link in social science. In that sense, causal sufficiency is more demanding in PCM applications.

A PCM advocate might acknowledge this point and reply that DAGs assuming causal sufficiency are a useful starting point in research. We can then consider, for which hypothesized links causal sufficiency is reasonable, and, subsequently, examine more complex models, in which it does not hold. So, violations of causal sufficiency do not suggest that PCM is not useful.

IV.b Pearl's equivalence claim

So, RCM advocates have good responses to PCM objections and *vice versa*. How can we get out of this impasse? Pearl claims to have proved the mathematical equivalence of RCM and PCM: any theorem in RCM has an equivalent theorem in PCM (Pearl 2012; Pearl, 2009, Chapter 7; Galles and Pearl, 1998; Glymour et al., 2016, p.126; Gelman, 2009). If true, this mathematical result has the potential to reconcile the two sides because it can show us which objections are justified given the shared axiomatic system of RCM and PCM. While Pearl's claim is not true, examining it will point us toward the source of the methodological disagreements.

Let us scrutinize Pearl's proof. He begins by outlining all assumptions embedded in his definition of a structural model, among which is the existence of POs (Pearl, 2009, Definition

7.1.4). Next, he presents a set of three axioms which are necessary and sufficient for deriving the definition of a structural model (Theorems 7.3.3 and 7.3.5). In other words, all properties which Pearl includes in his definition of a structural model (including POs) can be inferred from the axioms. So, the three basic axioms are the building blocks for more complex statements made within structural models: any claim involving POs can be reduced to a claim expressed in the axioms. Consequently, any result which Rubin derives using POs can be translated into PCM.

While I do not doubt Pearl's mathematical derivation, I do doubt whether Rubin will accept his definition of POs. The reason is that it implies that the *same* structural model applies to all units. Consider Pearl's definition of POs:

$$Y(do(D = 1), M(b)) \tag{4}$$

where Y is the outcome of interest, D is a binary treatment indicator, b are background factors and M is a model which connects the variables of interest. The do-operator $do(.)$ provides rules which we can use to calculate the effect of specific interventions within M, e.g., setting $D = 1$.[18] For Pearl, POs are inseparable from the model M linking b.

This feature of POs within PCM is important since it suggests that causal effects will be the same for two people with the same b. It is also reflected in one of Pearl's axioms: the axiom of composition (Pearl, 2009, p.229).[19] It implies that if two people share the same characteristics b in model M, they will have the same POs once we intervene to set $D = 1$ where these predicted POs will be the same as the actual POs. This means that $D = 1$ will have exactly the same effect on their POs. For example, if two people share the same education and health in Figure 1, changing their neighborhood will have exactly the same effect on their incomes.

Let us now see how POs in RCM compare to POs in PCM. If we stick to do-calculus, RCM POs are given by $Y_j(do(D = 1))$ where j denotes the unit of observation (Imbens and Rubin, 2015) and where I drop time indexing for brevity. In this definition, Rubin allows POs to differ across people, independently of the underlying structural model, so it is *not* always true that

[18] I will not discuss $do(.)$ in detail, since it helps causal identification, not discovery.

[19] More formally, composition states that the only variable changed *directly* by the intervention is the treatment indicator, given M. The descendants of the treatment indicator change but only *indirectly* as a result of the intervention.

(Markus, 2021):

$$\underbrace{Y(do(D=1), M(b_j))}_{POs\ in\ PCM} = \underbrace{Y_j(do(D=1))}_{POs\ in\ RCM}$$

Even if i and j share the same characteristics ($b_i = b_j$), they might not share the same M on RCM. So, the difference between Rubin and Pearl boils down to a disagreement about how POs are generated across observationally equivalent people. For Pearl, the variation between individuals can be fully captured by the variables in M whereas for Rubin this might not be the case. Rubin will reject the equivalence claim not simply because Pearl defines POs within a model but because the model imposes the same POs on two people with the same b and, therefore, the same causal effect. This distinction is important because Markus (2021) thinks that defining POs within a model is the issue. For me, this does not capture the essence of Rubin's response to Pearl.

This reasoning somewhat simplifies things. Both Pearl and Rubin allow for the effect of random shocks to affect POs. For Pearl, probabilistically independent random disturbances are allowed to affect POs independently of the other causes. However, for Rubin, they can affect both the POs *and* the causal effect of the treatment. Even if the POs for i and j are affected by equal random shocks, the causal effect might still be different for Rubin: there might be an additional random shock in RCM, affecting just the causal effect itself.

IV.c The *fundamental problem*

To appreciate the importance of the last point, we need to consider the *fundamental problem of causal inference*, i.e., if we can estimate the individual causal effect in an actual causation case. In PCM, this is possible: we can learn how much neighborhood changes the income for a particular individual (Pearl, 2009, §10). However, RCM sees the *fundamental problem* as the most important issue in causal inference because we never observe both POs.

So, RCM advocates will object to PCM because it mistakenly claims to resolve the *fundamental problem*. How can RCM advocates support this claim? One strategy would be to provide examples of individual causal effects in actual causation cases which PCM cannot han-

dle. Specifically, consider a single equation determining income:

$$Income_j = \beta Education_j + u_j \tag{5}$$

where u_j is a probabilistically independent error term and where causal sufficiency holds. In (5), we can clearly see that causal effects are individually invariant, i.e., β is the same for everybody. So, the incomes of two people with the same education will differ if and only if their error terms differ. This implication looks unreasonable. For example, there is substantial evidence that the returns to education differ across genders (Dougherty, 2005).

Thus, we might want to introduce such variation by modeling the true causal effect as $\beta_j = \beta + \epsilon_j$. In this formulation, ϵ_j does not enter directly into either education or income but only into the causal effect. RCM methods can still tell us something about the average treatment effect under relatively general conditions on the probability distribution of ϵ_j (Imbens and Wooldridge, 2009). Unfortunately, it is not obvious how we can add such unobserved randomness in the arrows of a DAG. We cannot treat u_j as a mediator since it is unobserved. In that sense, PCM allows us to introduce randomness into variables but not into causal effects. In situations, where randomness enters into the causal effects, PCM is unable to capture actual causation. In other words, RCM advocates are right in saying that PCM cannot always resolve the *fundamental problem*. Appendix A.b provides another example of such a situation.

IV.d The importance of the *fundamental problem*

So far, we have considered four clash points in the methodological debate between RCM and PCM: (i) RCM's inability to explore mechanisms, (ii) the importance of causal sufficiency in PCM, (iii) Pearl's equivalence claims and (iv) the *fundamental problem*. In my view, the disagreement about the *fundamental problem* is the most important one since it can explain why PCM and RCM advocates disagree about the other three issues. Let us see why this is the case. Because of the *fundamental problem*, RCM advocates believe that we can never fully recover the mechanism behind a causal effect in an actual causation case. So, they reject Pearl's definition of POs and his equivalence claim, i.e., (iii). In contrast, for PCM advocates, the *fundamental*

problem can be resolved by modeling the causal transmission mechanism in a DAG, and thus Pearl's equivalence claim makes sense.

Why do they disagree about mechanisms, i.e., (i)? Since PCM can resolve the *fundamental problem* via modeling the mechanism, there is no problem in using data from one individual to learn about the mechanism of another. On the other hand, given the *fundamental problem*, RCM postulates that it is hard enough to establish the existence of a single causal effect, let alone the whole mechanism, in which this causal effect is embodied. Similarly, RCM advocates think that given the nature of social science (Imbens, 2020) and that we can never learn about the causal transmission for a particular individual, it is extremely likely that in our data there is an individual, for whom there is a common cause to both variables in the causal link of interest. In other words, causal sufficiency will almost always fail, i.e., (ii). Such individual violations are less concerning for PCM advocates because they should cancel out in the aggregate model we use to resolve the *fundamental problem*.

V Traditional metaphysical theories

V.a Introduction

Although there are various methodological disagreements between PCM and RCM advocates, I showed that they boil down to a disagreement about the *fundamental problem of causal inference*, i.e., whether we can estimate individual causal effects in actual causation cases. Despite the importance of the *fundamental problem*, this reasoning does not tell us *why* PCM and RCM assume it has different solutions in the first place. For that purpose, we need to venture into the metaphysics of causation. This approach is justified because actual causation is a major topic in philosophy of causation (Russo and Williamson, 2011). In addition, we cannot explain why different methods work without talking about metaphysics,[20] i.e., '[m]etaphysics and methodology should go hand in hand' (Cartwright, 2007, Chapter 9). So, discussing the metaphysics of

[20]Even if one thinks that this is insufficient to link metaphysics and methods, there are two reasons for this link. First, linking a metaphysical theory with a particular method can elucidate some of the problems faced by that method. Cartwright (2010) discusses how the probabilistic theory sheds light on the external validity of RCM methods. Second, all causal inference methods make implicit metaphysical assumptions. Studying how methods operate helps us *explicate* these assumptions which remain hidden behind formal definitions (Hu, 2022).

causation is justified. In the remainder of this paper, I argue that PCM and RCM depend on two slightly different metaphysics of causation which explains their responses to the *fundamental problem*.

Unfortunately, I cannot base my argument on the traditional theories of causation because they cannot explain why causal inference based on RCM or PCM works. Consequently, they also cannot explain their methodological differences. We need an alternative theory. Before introducing such a theory, this section shows why the mainstream versions of the probabilistic, regularity, counterfactual, and interventionist theory will not do the job (Beebee et al., 2009).[21]

V.b Regularity theory

The regularity theory (Psillos, 2009, p.131) states that event c causes event e iff:

(i) Events c and e are contiguous in spacetime;

(ii) c precedes e temporally;

(iii) All events of type C are regularly followed by all events of type E.

In this definition, c and e refer to specific *events*, e.g., Alexei's neighborhood and income, whereas C and E refer to *event-types*, e.g., Kensington and being rich. Conditions (i) and (ii) ensure that c and e are two closely linked events with the correct temporal order. Condition (iii) introduces the idea of regularity between event-types where a regularity is a lawlike universality and not just a generalization over separate instances (Psillos, 2009). So, if we observe that everybody brought up in Kensington is rich, then bringing up Alexei in Kensington will make him rich.

Unfortunately, the regularity theory might incorrectly classify events linked by common causes as cases of causation. Consider the fact that on Earth all event-types *day* are followed by event-types *night*. By the regularity theory, there is nothing to stop us from saying that today will cause tonight to appear. This is clearly false since today and tonight have a common cause: earth rotation. Our theory of causation needs to rule out such regularities from a common cause.

We can easily deal with such counterexamples in PCM and RCM. In PCM we need causal sufficiency before concluding that today will cause tonight. We can try adding causal sufficiency

[21] I exclude Salmon's process theory because it seems less applicable to social sciences.

to the regularity theory but it is not obvious how it can be defined in terms of regularities. In RCM the (to)day-causing-(to)night example is ill-defined. Our POs are night and our treatment is day: $Night_i(Day = 1)$ and $Night_i(Day = 0)$. To conduct causal inference, we need to observe at least one unit i, on which we can give the treatment $Day = 0$. This is clearly impossible, given earth rotation. In other words, to implement treatment $Day = 0$, we need to stop earth rotation, but this means that we have acknowledged earth rotation as a common cause. So, the basic regularity theory cannot explain why causal inference techniques work because it cannot deal with common cause counterexamples which are not an issue on PCM and RCM.

V.c Counterfactual theory

This theory (Paul, 2009) states that C is a cause of E iff:

(i) The counterfactual 'if not C, then not E' is true.

The theory tries to reduce causation to counterfactuals which are taken as a primitive. How do we evaluate if a counterfactual is true? Lewis suggests using his similarity semantics: counterfactual $A \rightarrow B$ is true if it is true in the A-closest possible world. In other words, a world in which A and B are both true is closer to our actual world than a world in which A holds but B does not (Menzies and Beebee, 2020, §1.1). Equipped with this similarity semantics, we can evaluate counterfactuals for evidence of causation.

If we set questions about preemption (Hall and Paul, 2003) aside,[22] the main problem with using Lewis' theory to explain causal inference methods is that RCM and PCM do not take counterfactuals as a primitive. Counterfactuals in RCM only provide an alternative reading of the definition of POs and play no other role, e.g., we cannot examine the validity of our reading of POs as counterfactuals within PCM. Counterfactuals in PCM emerge when we try to identify causal effects within a graph and, crucially, rely on a causal modelling semantics rather than a similarity one. So, even if they are taken as a primitive in PCM, they follow

[22]While many believe that these cases show Lewis' counterfactual theory to be false (e.g. Reiss, 2009, p.22), this has been questioned recently (Northcott, 2021; Clarke, 2022). Here is a preemption case. Suppose we have two processes $C1$ and $C2$ that are both sufficient to cause E (Collins, 2004). If E is becoming a philosopher, then $C1$ and $C2$ can be two PhD offers I am about to receive. Just before $C1$ causes E, $C2$ causes E. In this case, the counterfactual 'if not $C2$, then not E' is not true because $C1$ would have caused E if $C2$ had not occurred. On the counterfactual theory, $C2$ is not E's cause, although intuitively we would like to say $C2$ caused E.

very different inference rules from Lewis' semantics. While Pearl has claimed the equivalence of Lewis' semantics with his own (Glymour et al., 2016, p.116), Briggs (2012) has argued that certain counterfactuals (including *modus ponens*) are evaluated differently in the two frameworks, meaning that there is a counterexample disproving Pearl's claim.

V.d Probabilistic theory

This theory states that C causes E iff:

(i) C precedes E temporally;

(ii) $Pr(E|C) > Pr(E)$;

(iii) There is no B occurring earlier than C such that $Pr(E|C\&B) = Pr(E|B)$.

Here C and E are event-types but they can be single events in alternative versions of the theory. Condition (ii) states that the cause must be raising the probability of the effect whereas condition (iii) ensures that we have not missed any common cause B which would render C and E conditionally uncorrelated. The probabilistic theory is not subject to the day-causing-night counterexample due to condition (iii).

There are many well-known issues with the probabilistic theory such as its inability to remain a fully reductive theory of causation in the face of Simpson's paradox (Hitchcock, 2021). In my view, the most obvious reason why the theory fails are counterexamples such as faithfulness failures discussed in Section II. Using our old example, conditional on occupation we get:

$$P(Wage|Gender\&Occupation) = P(Wage|Occupation)$$

The probabilistic theory implies that there is no causal effect of $Gender$ on $Wage$ but this is not necessarily true.

Can the probabilistic theory justify PCM? Pearl (2009, §1.3-1.4) acknowledges that he can either use the probabilistic theory *or* the structural approach to read DAGs causally (Section II). He chooses the latter approach partly because it does not rely on the probabilistic theory but also because it allows evaluating counterfactuals (Pearl, 2009, p.27). So, while it can ground parts of

PCM, the theory is clearly not Pearl's favored approach. Similarly, Holland (1986, §5.3) argues that the probabilistic theory is 'quite different' from RCM because (a) it imposes *no* restrictions on what can count as a cause in contrast to RCM's manipulability requirement and (b) it discusses probabilities on the aggregate, not the individual level. Even if we define probabilities on the individual level, it is not obvious how to aggregate them across subpopulations (Hitchcock, 2021, §2.6).

V.e Interventionist theory

On this theory (Woodward, 2016), C causes E iff:

(i) Intervening on C in the right way will result in a corresponding change to E.

The key notion here is 'intervention'. A good definition of 'intervention' will also tell us how to intervene 'in the right way'. Woodward (2016) differentiates two ways to define intervention: Pearl (2009, §3) understands an intervention relative to a structural model whereas Woodward's definition is model-less (Woodward and Hitchcock, 2003). This distinction matters because PCM and RCM understand the notion of intervention differently (Markus, 2021, p.448). While RCM considers empirical interventions in the real world, PCM studies interventions defined relative to structural models. So, our best shot at using the interventionist theory to explain causal inference is to examine if (i) Pearl's version of the interventionist theory[23] can explain PCM and if (ii) Woodward's version of the interventionist theory can explain RCM.

Consider (i). Unfortunately, Pearl's version is insufficient to do the job. We can evaluate the effect of interventions only *within* a causal model. In other words, we need to first discover a DAG and justify its causal reading before examining interventions. However, the interventionist theory does not tell us how to do this. In particular, Cartwright (2007, Part II) has cast doubts on proofs aimed at showing that one of the necessary assumptions in causal discovery algorithms (Causal Markov Condition) can be derived solely from the interventionist theory.

[23]Pearl did not personally offer an interventionist theory. He only provided a definition of intervention which can be used as a basis for an interventionist theory (Woodward, 2016). So, it is useful to call the interventionist theory based on Pearl's notion of intervention Pearl's interventionist theory of causation, so that we can distinguish it from other interventionist theories. More generally, the literature is inconclusive about the metaphysics of causation assumed by Pearl: while Woodward (2016) credits Pearl as advocating an interventionist theory, Papineau (2021) argues that we can only make sense of PCM via the structural theory discussed below.

What about RCM and Woodward's interventionist theory? Note that the interventionist theory is not reductive. As a result, it cannot explain *why* a causal effect found from RCM methods obtains. While the regularity theorist can explain it via empirical regularities and the probability theorist can explain it via (probabilistic) difference-making, the interventionist theorist can only say that we have intervened in the right way. Intuitively, this does not seem right because it is begging the question. For the interventionist theorist, if RCM has discovered a causal effect, it must be because intervening on the cause will produce the effect but this is exactly what we have done with our RCM method.

The standard response is to say that the interventionist theory might not be truly reductive but it can still be useful (Woodward, 2016). In our case, this response is unconvincing because usefulness is not a good metric in the realm of metaphysics. It might help us adjudicate between different means of transport or between different causal inference techniques, e.g., if PCM is more useful than RCM, we should prefer it. However, it does not provide a reason to prefer one metaphysical theory over another. This matters because below I will offer a reductive theory of causation. So, even if the interventionist theory can explain RCM, it is not the best option from a metaphysical standpoint.

VI Structural equations theory of causation

The previous section suggests that metaphysics can help us resolve the methodological disagreements. However, it also suggests that we need an alternative metaphysical theory to achieve this aim because the standard theories cannot help us. Inspired by Papineau (2021, 2022), I argue for the *structural equations theory of causation*. It reduces causation to 'underlying structural equations with probabilistic exogenous terms' (Papineau, 2021, p.248). In this theory, there is a system of structural equations behind every causal mechanism we might want to study. In practice, we only observe bits of these structural equations. We probably do not observe all variables from the true system. This creates the need for causal inference techniques. In other words, if we would like to learn causal effects and we do not observe the true causal mechanism, we need to decide on the best way to leverage the available information which is done by causal

inference techniques. Unfortunately, I show that Papineau's original version is *insufficient* to explain why **P**CM works because it does not impose enough restrictions on the functional form of the structural equations. I then offer a slightly modified version of the theory which deals with the insufficiency problem and also explains why **R**CM works.

We can gain insight into the structural equations theory by getting a bit more formal. There is a set of *true* K^* equations that govern all mechanisms in nature and society. The k-th structural equation for the outcome variable y_k^* is:

$$y_k^* = f_k^*(pa_k^*) + u_k^* \tag{6}$$

where pa_k^* is the full set of parents of y_k^* and u_k^* is a probabilistically independent error term which could always be generated by chancy quantum effects (Papineau, 2021, §21). Superscript * indicates *true* variables in contrast to *observed* ones without superscripts. So, (6) states that the variables on the right-hand side determine the outcome variable via the functional form $f_k^*(.)$. The right-hand side only contains variables that have appeared in one of the previous $k-1$ equations and exogenous error terms. This means that the system of equations does not contain cycles, i.e., it is a DAG. Definition 1 which is closely linked with Simon (1953) summarises the theory. Conditions (ii) and (iii) ensure that there is a system of R equations, containing a directed causal path from C to E.

Definition 1 (Structural Equations Theory)**.** *C causes E iff:*

(i) There is a recursive mechanism consisting of K^* equations of the form (6);

(ii) Variables C and E are connected via $1 \leq R \leq K^*$ equations;

(iii) E is an ancestor of C within this system of R equations.

Does Papineau's theory in Definition 1 differ from Hausman's **CP** principle (1998)? One crucial difference is that Hausman's theory is not reductive (Hitchcock, 2000). In contrast, the structural equations in Papineau's theory represent lawlike regularities that actually appear in nature. This establishes a connection with the (reductive) regularity theory of causation (Papineau, 2021, p.31).

My version of the theory, however, differs in a crucial respect from Papineau's. The difference is in the restrictions on the functional form of the structural equations. Compare:

$$\underbrace{y_k^* = f_k^*(pa_k^*, u_k^*)}_{Papineau's\ eq.\ (22)} \qquad vs. \qquad \underbrace{y_k^* = f_k^*(pa_k^*) + u_k^*}_{My\ eq.\ (6)} \tag{7}$$

The u_k^*-s are restricted to enter additively into the outcome variable in our formulation. To see why this matters, it is necessary to unpack Papineau's argument for the structural theory (2021, §13). As in Definition 1, he assumes a system of K^* equations. Importantly, each structural equation has an unobserved error that is probabilistically independent from other causes. Papineau then claims that this condition is sufficient to always differentiate the true causal structure from alternatives. This result is crucial for Papineau's argument for two reasons. We can use it to explain why PCM-based causal inference works and it also confirms his structural theory of causation.

Why do functional forms matter for Papineau's argument? Consider his main claim:

Claim 1. *If the errors in each true structural equation are independent from other causes, then we can always differentiate the true causal system from alternatives.*

Unfortunately, this claim is not always correct when we work with Papineau's functional form $y_k = f_k(pa_k, u_k)$. In Appendix A.c, I provide a formal example of such a situation. However, the example does not tell us *why* Claim 1 is false. I believe that it fails because Papineau needs to prove two separate claims before establishing Claim 1. The first is:

Claim 2. *If the errors in each true structural equation are independent of other causes, (*) then the errors* ***enter*** *the outcome variable independently from other causes.*

According to Claim 2, the probabilistic independence of our errors implies that they actually enter the outcome variable independently from these causes.[24] Then, he needs to establish:

Claim 3. *(*) If the error terms* ***enter*** *the outcome variable independently from other causes, then we can always differentiate the true causal system from alternatives.*

[24]Consider $y_k = (pa_{1,k} + pa_{2,k})u_k$. Here u_k is probabilistically independent of both parents but does not enter into y_k independently from them. Appendix A.c discusses this example further.

If Claims 2 and 3 are both true, then Claim 1 is also true. However, the example in Appendix A disproves Claim 2: Papineau's theory fails because it incorrectly assumes the validity of Claim 2 under his functional form. Luckily, Claim 2 is true with the functional form I have suggested because the error terms are constrained to enter additively into the outcome.[25] So, if Claim 3 is correct, then my version of the structural theory implies the validity of Claim 1, meaning that it successfully reduces causation to structural equations.

Having established this point, we can show how the structural theory explains causal inference techniques. The K^* equations and the true variables define the complete causal mechanism. In practice, however, we are often interested in modeling one particular outcome such as income in Figure 1. We usually observe only a subset of all variables and equations from the true mechanism. As a result, we can at most recover K equations of the form:

$$y_k = f_k(pa_k) + u_k \tag{8}$$

where we are interested in modelling y_k and where pa_k is not necessarily a subset of the true pa_k^*, e.g., we might have not observe the l-th true parent $pa_{k,l}^*$ but we might still include one of $pa_{k,l}^*$'s *ancestors*. If we are willing to make some additional assumptions such as causal sufficiency, we can now conduct causal inference. So, we can infer causal effects, even when we do not observe the full true system.

In that sense, Papineau's reduction of causation to structural equations remains very faithful to PCM. It makes the structural causal models that Pearl uses for causal readings of DAGs into the metaphysics of causation. We can summarise these ideas by formalizing Definition 1. There is a true mechanism of R equations which connects outcome y_k^* and treatment x_k^* as in $y_k^* = M^*(b_k^*, x_k^*)$ where b_k^* are all the relevant ancestors of y_k^* which could be mediators or background factors. However, we are trying to model that mechanism using observed variables $y_k = M(b_k, x_k)$ where b_k is a proxy of the true b_k^* and x_k proxies[26] x_k^*. While $M^*(.)$ refers to the true *mechanism*, $M(.)$ refers to the *model* we use for causal inference.

We can now see how causal inference works in practice. Let the true pathway for the trans-

[25] Our formulation can probably be relaxed to $y_k^* = f_k^*(pa_k^*) + g_k^*(u_k^*)$ where $g_k^*(.)$ is a measurable function.

[26] This suggests that we do not have to change the actual treatment x_k^*, in which we are interested. Instead, we might use as treatment a proxy x_k, on which it is easier to intervene.

mission of pain be $M^*(b_k^*, x_k^*)$ where x_k^* is damage applied to my hand. The mechanism contains all the variables (neurons) along the ascending spinal tract. However, as an observer, I can only see the initial damage to my hand (x_k) and the degree of pain I feel (y_k). Thus, I can conclude that more damage to my hand causes more pain, even though I do not observe most of the true mechanism M^* which I approximate with model M.

From an RCM standpoint, there are two issues with this theory. First, an RCM advocate might object that the structural equations theory already gives raise to a particular solution to the *fundamental problem*. Similarly to PCM, the theory implies that we can estimate individual causal effects if our model approximates well the true mechanism. Given that individuals share the same mechanism, we can use data from different people to obtain our model. It also implies that if two people share exactly the same characteristics, they will respond in exactly the same way to a treatment.

Second, the structural theory seems unable to capture the evolution of causal mechanisms over time. It assumes that the system of structural equations governing my behavior is the same as the system governing the behavior of a person in another era. However, it is unlikely that coding skills mattered for incomes in the Middle Ages but they do matter for incomes nowadays. This is not a trivial point because it shows that systems of structural equations are might change over time. In other words, different factors get different weighting at different times. In practice, this assumption is often unproblematic. Probably, the same model determines my income now and three years ago. Nevertheless, it is unclear why we are justified in assuming time-invariant structural equations in our metaphysics.

So, the standard structural equations theory which we can call *Pearl's* structural equations theory (PSET) would probably be unacceptable to RCM advocates. Thus, we need *Rubin's* structural equations theory (RSET) to justify RCM. Luckily, we only need to tweak PSET slightly to get RSET. We simply make the structural equations from PSET individual-specific. To that aim, we denote individuals as j, so that the true causal mechanism for individual j is given by K_j^* equations of the form:

$$y_{k,j}^* = f_{k,j}^*(pa_{k,j}^*) + u_{k,j}^* \tag{9}$$

We allow the functions $f^*_{k,j}(.)$ to be individual-specific for j and the number of equations K^*_j to differ across individuals. These are the only differences with PSET. In particular, if we add j-notation to PSET, it postulates $f^*_k(pa^*_{k,j})$ whereas RSET postulates individual-specific $f^*_{k,\boldsymbol{j}}(pa^*_{k,j})$. The effect of a particular treatment is given by the individual-specific true mechanism $y^*_{k,j} = M^*_j(b^*_j, x^*_j)$. For causal inference, we need to approximate the true mechanism with an individual-*specific* model $y_{k,j} = M_j(b_j, x_j)$ on RSET. In contrast, PSET postulates the individual-*invariant* mechanism $y^*_{k,j} = M^*(b^*_j, x^*_j)$ and model $y_{k,j} = M(b_j, x_j)$ which have no individual subscripts.

Why would RCM advocates prefer RSET? Causal effects are by definition individual-specific on RSET, even for people with the same covariates. So, we cannot hope to calculate the causal effect for a given individual by using information from other individuals, i.e., RSET implies the *fundamental problem*. Moreover, RSET allows us to capture changes in the causal mechanism through time. While the $M^*_j(.)$-s for people in the Middle Ages might have contained no place for an effect of coding skills on income, these people passed away and made away for people nowadays whose $M^*_j(.)$-s do contain a place for coding skills. This is also in line with the time-indexing of POs in PCM (Section II).

Markus (2021) suggests that RCM advocates might not like defining POs relative to a model, meaning that they may reject RSET. However, Markus' argument assumes an individual-invariant model. Relaxing this assumption makes models much more acceptable because they are akin to a data-generating process. Statisticians often use such processes when studying the theoretical properties of different estimators.[27] So, if assuming a true but inaccessible data-generating process is acceptable to RCM advocates, then assuming individual-specific inaccessible structural equations should also be acceptable.

How can RSET and PSET explain the methodological differences between RCM and PCM? We can see the different responses to the *fundamental problem*. On PSET, individuals share the same mechanisms, so we can use your mechanism as a guide to mine. Once we estimate the parameters of our model we can also derive the causal effects in actual causation cases. In contrast, since the mechanisms might be very different across people on RSET, this is not an option. This also explains why RCM advocates do not generally examine mechanisms. There

[27]For example, we might investigate how well the ordinary least squares estimator performs under a data-generating process of a linear model with an omitted variable using asymptotic theory (Wooldridge, 2010, §4.3)

may or may not be a shared mechanism: we simply do not know in advance. For this reason, RCM methodologists develop techniques aimed at making as few assumptions as possible on these mechanisms. Moreover, because we can easily think of violations of causal sufficiency for a particular individual given RSET, RCM advocates are skeptical of this assumption. In contrast, individual violations of causal sufficiency are less problematic on PSET.

VII Conclusion

In this paper, I argued that the methodological differences between RCM and PCM can be explained by their different views of the metaphysics of causation. RCM and PCM advocates disagree about many things such as the importance of mechanisms. However, these disagreements boil down to a more basic disagreement about the *fundamental problem of causal inference*. Using this insight, I suggested that RCM and PCM both rely on a reduction of causation to structural equations but these equations vary across individuals on RCM. This allow us to see why they understand the *fundamental problem* differently and, consequently, why they disagree about other methodological issues.

These considerations raise the question if the metaphysical theory behind RCM or the one behind PCM is superior. In my view, there are three reasons why the RCM version is superior. First, there is no obvious metaphysical reason to impose individual-invariant mechanisms. Second, the RCM version allows us to capture the variation in the structural equations over time. As people pass away, mechanisms can change. Third, the RCM version is better suited to account for how we perceive the world. In practice, on any causal path, there are many unobservable mediators. For this reason, the data will seem to us *as if* it comes from the RCM metaphysics, even if it does not. This is a very brief answer to a much more complex question that I hope to explore in future work.

References

Andersen, H. (2013), 'When to expect violations of causal faithfulness and why it matters', *Philosophy of Science* **80**(5), 672–683.

Athey, S. and Imbens, G. W. (2017), 'The state of applied econometrics: Causality and policy evaluation', *Journal of Economic Perspectives* **31**(2), 3–32.

Baicker, K. and Finkelstein, A. (2011), 'The effects of Medicaid coverage—learning from the Oregon experiment', *The New England Journal of Medicine* **365**(8), 683–685.

Beebee, H., Hitchcock, C. and Menzies, P. (2009), *The Oxford Handbook of Causation*, Oxford University Press.

Bertrand, M. and Mullainathan, S. (2004), 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination', *American Economic Review* **94**(4), 991–1013.

Blau, F. D. and Kahn, L. M. (2017), 'The gender wage gap: Extent, trends, and explanations', *Journal of Economic Literature* **55**(3), 789–865.

Briggs, R. (2012), 'Interventionist counterfactuals', *Philosophical studies* **160**(1), 139–166.

Card, D. (1999), The Causal Effect of Education on Earnings, *in* O. C. Ashenfelter and D. Card, eds, 'Handbook of Labor Economics', Vol. 3, Elsevier, pp. 1801–1863.

Card, D., Cardoso, A. R. and Kline, P. (2016), 'Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women', *The Quarterly Journal of Economics* **131**(2), 633–686.

Cartwright, N. (2007), *Hunting causes and using them: Approaches in philosophy and economics*, Cambridge University Press.

Cartwright, N. (2010), 'What are randomised controlled trials good for?', *Philosophical studies* **147**(1), 59–70.

Cerulli, G. (2015), Methods based on selection on unobservables, *in* 'Econometric Evaluation of Socio-Economic Programs', Springer, pp. 161–227.

Chetty, R. and Hendren, N. (2018), 'The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects', *The Quarterly Journal of Economics* **133**(3), 1107–1162.

Chetty, R., Hendren, N. and Katz, L. F. (2016), 'The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment', *American Economic Review* **106**(4), 855–902.

Chyn, E. and Katz, L. F. (2021), 'Neighborhoods matter: Assessing the evidence for place effects', *Journal of Economic Perspectives* **35**(4), 197–222.

Clarke, C. (2022), 'Why your causal intuitions are corrupt: Intermediate and enabling variables', *Erkenntnis* .

Collins, J. (2004), 'Preemptive prevention', *The Journal of Philosophy* **97**(4), 223–234.

Dale, S. B. and Krueger, A. B. (2002), 'Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables', *The Quarterly Journal of Economics* **117**(4), 1491–1527.

Deaton, A. and Cartwright, N. (2018), 'Understanding and misunderstanding randomized controlled trials', *Social Science & Medicine* **210**, 2–21.

DiTraglia, F. J., Garcia-Jimeno, C., O'Keeffe-O'Donovan, R. and Sanchez-Becerra, A. (2020), 'Identifying causal effects in experiments with spillovers and non-compliance', *arXiv preprint No. 2011.07051* .

Dougherty, C. (2005), 'Why are the returns to schooling higher for women than for men?', *Journal of Human Resources* **40**(4), 969–988.

Galles, D. and Pearl, J. (1998), 'An axiomatic characterization of causal counterfactuals', *Foundations of Science* **3**(1), 151–182.

Gelman, A. (2009), 'More on Pearl/Rubin, this time focusing on a couple of points', https://statmodeling.stat.columbia.edu/2009/07/09/more_on_pearlru/.

Glymour, M., Pearl, J. and Jewell, N. P. (2016), *Causal inference in statistics: A primer*, John Wiley and Sons.

Goldin, C., Katz, L. F. and Kuziemko, I. (2006), 'The homecoming of American college women: The reversal of the college gender gap', *Journal of Economic Perspectives* **20**(4), 133–156.

Hall, N. and Paul, L. A. (2003), Causation and preemption, *in* P. Clark and K. Hawley, eds, 'Philosophy of Science Today', Oxford University Press.

Hausman, D. M. (1998), *Causal Asymmetries*, Cambridge University Press.

Heckman, J. J. and Pinto, R. (2022), 'Causality and econometrics', *NBER Working papers No. 29787* .

Hitchcock, C. (2000), 'Review: Causal asymmetries by daniel m. hausman', *British Journal for the Philosophy of Science* **51**(1), 175–180.

Hitchcock, C. (2021), Probabilistic Causation, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Spring 2021 edn, Metaphysics Research Lab, Stanford University.

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**(396), 945–960.

Hu, L. (2022), 'Causation in the social world', *PhD Dissertation* .
URL: *https://dash.harvard.edu/handle/1/37372043*

Imbens, G. W. (2020), 'Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics', *Journal of Economic Literature* **58**(4), 1129–79.

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**(1), 5–86.

Kauppinen, T. M. (2008), 'Schools as mediators of neighbourhood effects on choice between vocational and academic tracks of secondary education in Helsinki', *European Sociological Review* **24**(3), 379–391.

Kling, J. R., Ludwig, J. and Katz, L. F. (2005), 'Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment', *The Quarterly Journal of Economics* **120**(1), 87–130.

Kohler-Hausmann, I. (2018), 'Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination', *Nw. UL Rev.* **113**, 1163.

LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review* **76**(4), 604–620.

Lewbel, A. (2019), 'The identification zoo: Meanings of identification in econometrics', *Journal of Economic Literature* **57**(4), 835–903.

Ludwig, V. and Brüderl, J. (2018), 'Is there a male marital wage premium? new evidence from the united states', *American Sociological Review* **83**(4), 744–770.

Markus, K. A. (2021), 'Causal effects and counterfactual conditionals: contrasting Rubin, Lewis and Pearl', *Economics & Philosophy* **37**(3), 441–461.

Menzies, P. and Beebee, H. (2020), Counterfactual Theories of Causation, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2020 edn, Metaphysics Research Lab, Stanford University.

Northcott, R. (2021), 'Pre-emption cases may support, not undermine, the counterfactual theory of causation', *Synthese* **198**(1), 537–555.

Ogburn, E. L. and VanderWeele, T. J. (2014), 'Causal diagrams for interference', *Statistical science* **29**(4), 559–578.

Papineau, D. (1986), 'Causal factors, causal inference, causal explanation II', *Proceedings of the Aristotelian Society, Supplementary Volumes* **60**, 114–136.

Papineau, D. (2021), 'The Causal Structure of Reality', *Working paper* .
URL: *https://www.davidpapineau.co.uk/articles-online.html*

Papineau, D. (2022), 'The Statistical Nature of Causation', *The Monist* **105**(2), 247–275.

Paul, L. (2009), Counterfactual Theories, *in* H. Beebee, C. Hitchcock and P. Menzies, eds, 'The Oxford Handbook of Causation', Oxford University Press, p. 158–184.

Pearl, J. (2009), *Causality*, 2 edn, Cambridge University Press.

Pearl, J. (2012), 'Summary of my views on the relationships between the potential-outcome (PO) and Structural Causal Models (SCM) frameworks.'.
URL: *http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/ [Accessed: 23 Nov 2022]*

Psillos, S. (2009), Regularity Theories, *in* H. Beebee, C. Hitchcock and P. Menzies, eds, 'The Oxford Handbook of Causation', Oxford University Press, p. 131–157.

Reiss, J. (2009), 'Causation in the social sciences: Evidence, inference, and purpose', *Philosophy of the Social Sciences* **39**(1), 20–40.

Robert, S. A. (1999), 'Socioeconomic position and health: the independent contribution of community socioeconomic context', *Annual Review of Sociology* **25**, 489–516.

Russo, F. and Williamson, J. (2011), 'Generic versus single-case causality: the case of autopsy', *European Journal for Philosophy of Science* **1**(1), 47–69.

Sampson, R. J., Morenoff, J. D. and Gannon-Rowley, T. (2002), 'Assessing "neighborhood effects": Social processes and new directions in research', *Annual Review of Sociology* **28**, 443–478.

Scheines, R. (1997), 'An introduction to causal inference', *CMU Philosophy Working Paper* .
URL: *https://kilthub.cmu.edu/articles/journal-contribution/An-Introduction-to-Causal-Inference/6490904/1*

Simon, H. (1953), Causal Ordering and Identifiability, *in* W. C. Hood and T. C. Koopmans, eds, 'Studies in Econometric Method: Cowles Commission for Research in Economics', John Wiley & Sons, pp. 49–74.

Spirtes, P., Glymour, C. N. and Scheines, R. (2009), *Causation, prediction, and search*, 2nd edn, MIT press.

Spohn, W. (2009), Bayesian Nets Are All There Is to Causal Dependence, *in* W. Spohn, ed., 'Causation, Coherence, and Concepts: A Collection of Essays', Springer, pp. 99–111.

Stern, R. and Eva, B. (2022), 'Antireductionist Interventionism', *Forthcoming in The British Journal for the Philosophy of Science* .

Vagni, G. and Breen, R. (2021), 'Earnings and Income Penalties for Motherhood: Estimates for British Women Using the Individual Synthetic Control Method', *European Sociological Review* **37**(5), 834–848.

Weinberger, N. (2022), 'Comparing Rubin and Pearl's causal modelling frameworks: a commentary on Markus', *Economics & Philosophy* pp. 1–9.

Woodward, J. (2016), Causation and Manipulability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2016 edn, Metaphysics Research Lab, Stanford University.

Woodward, J. and Hitchcock, C. (2003), 'Explanatory generalizations, part i: A counterfactual account', *Noûs* **37**(1), 1–24.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Zhang, J. and Spirtes, P. (2008), 'Detection of unfaithfulness and robust causal inference', *Minds and Machines* **18**(2), 239–271.

Zhang, J. and Spirtes, P. (2016), 'The three faces of faithfulness', *Synthese* **193**(4), 1011–1027.

A Technical Appendix

A.a Graphical illustration of a faithfulness failure

Recall that in the example in Section II we found that conditional on occupation there is no correlation between gender and income. Figure 2 illustrates this example graphically. We asked if the absence of correlation between gender and income means that gender has no causal effect on income conditional on occupation. Unfortunately, this is not necessarily true. It is possible that gender affects income through two separate paths whose effects exactly offset each other. In Figure 2, we can see that gender *positively* affects income via education. This is because women are more likely to attend university, i.e., $G \xrightarrow{+} E$ (Goldin et al., 2006), and education increases earnings, i.e., $E \xrightarrow{+} I$ (Card, 1999). On the other hand, women's lower wage negotiating ability directly lowers their earnings, i.e., $G \xrightarrow{-} I$ (Card et al., 2016). Suppose that the direct *negative* effect on income *exactly* cancels out the *positive* effect mediated by education. So, there will be zero conditional correlation between gender and income. But this is a faithfulness failure because gender still affects income.

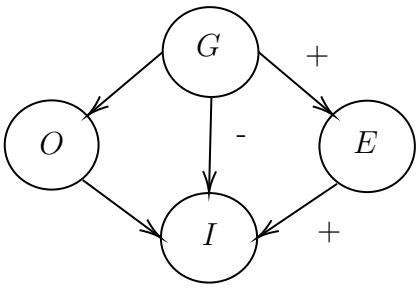

Figure 2: Faithfulness failure

A.b PCM and Actual Causation

This subsection provides another actual causation case that PCM cannot handle. Consider again the single equation determining income:

$$Income_j = \beta Education_j + u_j \tag{10}$$

where u_j is a probabilistically independent error term and where causal sufficiency holds. As noted above, there is substantial evidence that the returns to education differ across people. For example, they might differ across genders, even if gender has no independent effects on either income or education.[28] Crucially, this will not violate causal sufficiency because gender will not be a common cause of income or education. It will only affect the causal effect itself, not the treatment or the outcome variables. While we *can* estimate β in this case, our estimate will not be capturing the true causal effect contrary to Pearl's claim that we can get causal effects in cases of actual causation. It will only provide a summary of the returns to education across many different people. In RCM we generally cannot estimate individual effects as well. However, there is an explicit recognition that causal effects are individually varying (given the definition of POs) which is not the case in PCM.

Here PCM advocates might object that PCM can still allow β_j to vary across individuals by adding mediators. If we add gender as a mediator, we can model j's causal effect as:

$$\beta_j = \beta + \alpha_1 Gender_j + \alpha_2 Mediator_j \tag{11}$$

where β is the direct effect of education on wages. $Mediator_j$ is a second mediator and α_1 and α_2 give the contribution of the two mediators to β_j. However, this formulation raises more questions than it answers. Given that the effect of education on income differs across individuals, a mediator might have a different effect on different individuals. Education might affect earnings through a path containing two mediators and not just one. Its effect might be mediated through gender and marriage which is known to have contrasting effects on men (Ludwig and Brüderl, 2018) and women (Vagni and Breen, 2021). This reasoning suggests that at some point we must draw the line and say that there are no more mediators. Once this is done, the causal effect which we have estimated will be considered the same across all individuals. Drawing such a line will be necessarily arbitrary. It is not obvious when it will be reasonable to say that there are no more mediators. While such a threshold might exist, it is up to PCM advocates to provide a reasonable criterion for determining it.

[28] While this situation is rather implausible empirically, it is theoretically possible.

A.c Functional form objection to Papineau

This appendix provides a formal example that disproves a crucial assumption in Papineau's structural theory of causation (2021; 2022). More formally, Papineau's claim is:

Claim 4. *If the errors in each true structural equation are independent of other causes, then we can always differentiate the true causal system from alternatives irrespectively of the functional form.*

Unfortunately, the consequent does not always follow from the antecedent. That is, there exist functional forms under which we cannot distinguish the true and the alternative systems, even if the errors are independent. I proceed by proposing such a functional form which is the basis for the counterexample. Figure 3 gives the general set-up I consider.

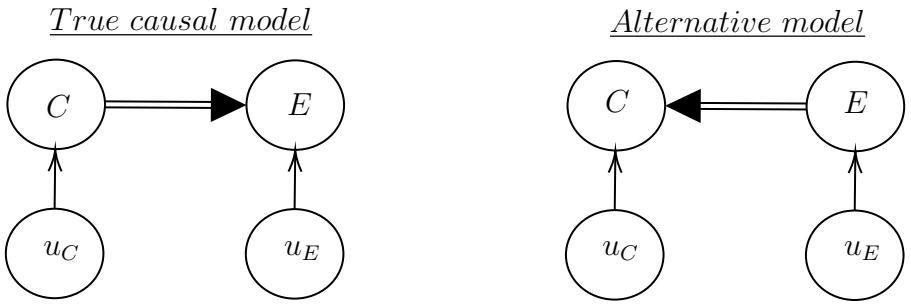

Figure 3: True and alternative DAG

We also assume (i) $E[u_c] = E[u_E] = 0$, (ii) $E[u_c^2] = Var(u_c) \neq 0 \neq Var(u_E) = E[u_E^2]$ and (iii) $E[u_c u_E] = Cov(u_c, u_E) = 0$. I will provide two functional form examples. Both functional forms are allowed by Papineau's formulation but only one of them allows us to distinguish the true causal system from alternatives.

A.c.1 Linearity

If we assume linearity, the DAGs reduce to one of Papineau's examples (2021, eq. (18), (20)). The linearity assumption is allowed by both Papineau's functional form and by our functional

form, as postulated in (7). The corresponding structural equations are:

True causal model	*Alternative model*
$C = u_C$	$C = \frac{1}{\beta}E + u_C^*$
$E = \beta C + u_E$	$E = u_E^*$

By substituting the true causal model into the alternative, we get:

$$u_E^* = \beta C + u_E$$
$$u_C^* = u_C - \frac{1}{\beta}E = u_C - \frac{1}{\beta}(\beta C + u_E) = -\frac{u_E}{\beta}$$

Then, we calculate:

$$Cov(E, u_C^*) = -\frac{E[u_E^2]}{\beta} \neq 0$$

Note two things. First, as expected on the structural theory, in the true model there is extra variation entering into E independent from its cause. Second, this is not the case in the alternative model where $Cov(E, u_C^*) \neq 0$. So, by Papineau's structural theory, we have correctly identified the true causal model.

A.c.2 Nonlinearities

Let us take a functional form allowed by Papineau's theory but not ours (recall eq. 7 above):

True causal model	*Alternative model*
$C = u_C$	$C = (\alpha + E)u_C^*$
$E = (\beta + C)u_E$	$E = u_E^*$

We simply use $f(pa_k, u_k) = (\beta + pa_k)u_k$ instead of the linear $f(pa_k, u_k) = \beta pa_k + u_k$. If $Cov(u_C^*, E) = 0$, we will be unable to distinguish the true from the alternative model. The reason is that in both cases we can separate the outcome variable into causes and independent

errors.

Let us evaluate this idea by following the same procedure as above:

$$u_E^* = (\beta + u_C)u_E$$
$$u_C^* = \frac{C}{\alpha + E} = \frac{u_C}{\alpha + (\beta + C)u_E}$$

Note that $Cov(\alpha + E, u_C^*) = E[(\alpha + E)u_C^*] = E[Eu_C^*]$. So, we have:

$$E[(\alpha + E)u_C^*] = E[u_C] = 0 \implies Cov(E, u_C^*) = 0$$

Similarly, $Cov((\alpha + u_E^*), u_C^*) = E[u_E^* u_C^*] = Cov(u_E^*, u_C^*) = 0$. In both the true and the alternative system, we can reduce the outcome variable into its causes and independent errors. So, the structural equations theory does not tell us which is the correct causal structure.