Norms for a good theory of causation[*]

Tzvetan Moev

1 Introduction

What is a good theory of causation? Philosophers give contrasting answers to this question. Take Jim Woodward, for example. He thinks that the interventionist theory of causation is a good theory because it is a useful theory (Woodward, 2014*b*). He also thinks that the counterfactual theory is *not* a good theory because it is *not* useful (Woodward, 2003, p.30). Compare these claims with Paul and Hall (2013). They consider the counterfactual theory a good theory because it is reductive. They also do *not* consider the interventionist theory a good theory because it is *not* reductive (Paul and Hall, 2013, p.26). Why do philosophers disagree about what makes a good theory of causation? How should we adjudicate between two different theories?

In this paper, I argue that different philosophers subscribe to different *norms* about what makes a good theory of causation. In epistemology, epistemic norms describe the rules that determine if (e.g.) an assertion is justified. If I make an assertion in accordance with your favorite epistemic norms, you ought to accept it. I contend that the same is true when adjudicating between theories of causation. Instead of *epistemic* norms, philosophers use *selection* norms for theories of causation. Thus, some philosophers say that we ought to prefer practical theories whereas others say that we ought to prefer

[*] Thanks to Madelaine Angelova-Elchinova, Alexander Bird, Jacob Stegenga, Rueben Stern, Binjie Zou, and members of Bulgarian Society for Analytic Philosophy (BSAP) for comments. Paper is currently updated in light of feedback from presentations at Duke Causation Group, PSA Around the World, and Duke Grad Students WIP seminar.

reductive theories. It seems that different philosophers subscribe to different selection norms. As a result, they also disagree about the value of different theories.

Let us consider two sets of selection norms for theories of causation used in philosophy. One set of such norms is offered by a group of philosophers that I refer to as *reductivists* such as Paul and Hall (2013). For them, a theory of causation is a good theory if it i) reduces causation to something more ontologically fundamental and ii) matches our causal judgment in a large set of tricky cases such as late preemption. For reductivists, the combination of selection norms i) and ii) determines whether we should prefer one theory of causation over another. Another set of selection norms is offered by *interventionists* such as Woodward (2003). They subscribe to what I call the *Functional Value Norm*: the greater the functional value of a theory, the better the theory is. Functional value roughly denotes how useful or practical a theory is for achieving some objectives that we care about. So, the *Functional Value Norm* tells us to prefer theories with greater functional value.

Which is the correct set of norms? I argue that while the original *Functional Value Norm* needs to be slightly revised, it is the most fundamental selection norm and thus is more important than other norms. So, we should use the *Functional Value Norm* when choosing between different theories. The rest of this paper provides support for this position. I begin by explaining what I mean by selection norms for theories of causation (§2). Next, I discuss the selection norms used by reductivists (§3) and interventionists (§4). I then argue why the *Functional Value Norm* needs a revision and how this should be achieved (§5). I provide two positive arguments in its favor (§6) before replying to some objections (§7).

2 Norms and theories of causation

Since we will be talking about norms a lot, it is important to clarify what I will *not* discuss. A lot of philosophers are interested in the effect of what we consider as normal on our causal judgments (Knobe and Fraser, 2008; Hitchcock and Knobe, 2009). Their work suggests that we are more likely to identify a factor as 'the cause' of an event if it violates a norm or, more generally, is considered as abnormal. While this finding raises various important questions (Henne, 2023, §2.1), we will not explore them. Why? Because I am not interested in how norms affect causal judgments but instead in how they (should) guide our choice between theories of causation. This latter question has both a descriptive and a normative dimension. The descriptive dimension is about the actual norms that different philosophers use to identify good theories of causation whereas the normative dimension is about which set of norms they *should* be using. I hope to shed light on both dimensions in this paper.

My discussion of norms is instead closer to how norms are discussed in epistemology (Littlejohn, 2013).[1] Epistemic norms tell us what we can legitimately believe, assert, etc. They provide the rules that we ought to follow when believing, asserting, etc. One famous example is Williamson's knowledge norm of assertion (2000, p.243):

> One ought to assert p only if one knows p

This norm tells us that knowing p is a necessary condition for asserting p. Unless I know that Joe Biden is the president of the US, I cannot assert it. Williamson's knowledge norm also allows us to pronounce judgments on the assertions made by others. Whenever you make an assertion, I can always ask 'How do you know this?' (Williamson, 2000, p.251). If you fail to provide a good answer, I am under no obligation to accept your assertion. For example, you might assert that I am wasting my time by buying a lottery ticket (Pritchard, 2013). I can then ask 'How do you know?' and you might answer that my ticket will lose. Assuming that I am participating in a fair lottery, I can then respond that your assertion is not justified: my ticket has a non-zero chance of winning and so you cannot know that it will definitely lose.

My contention in this paper is that there are norms that function in a similar way when philosophers *select* among different theories of causation. If a theory respects your favorite *selection* norms, you ought to accept it. So, selection norms help us determine

[1] Normativity is also a major topic in ethics and aesthetics (among other fields) but I found the discussion in epistemology to be the most insightful for my purposes.

if a theory is a good theory in the same way in which Williamson's knowledge norm helps us determine if an assertion is a good assertion. Unfortunately, selection norms are often not stated explicitly by philosophers working on causation. However, they seem to be implicitly assumed. As we saw above, some philosophers think that we ought to reject impractical theories. Others think that we ought to reject non-reductive theories. For this reason, I assume that the best way to make sense of the different rules used by philosophers to rank theories of causation is to grant the existence of such selection norms.[2]

At this point, one might ask why we refer to the rules philosophers use for selecting theories of causation as *norms* instead of *selection criteria* or *theoretical commitments* or a related concept. In principle, all of these terms can work and the subsequent discussion does not hinge on which term we use. However, norms have several features that make them particularly apt. First of all, norm is the state-of-the-art term that philosophers use to explain how we make judgments and here we are interested in the judgments that scholars make about theories of causation. Given the complexity of normativity, this term also suggests that any set of norms is potentially open to debate. In contrast, *selection criteria* imply that there might be a uniquely correct set of rules. On the other hand, *theoretical commitments* suggest implicit assumptions, to which one gets committed when adopting a particular theory of causation (Horwich, 1991).[3] However, norms (as I intend them) can be outlined *before* one adopts a particular theory and so there is a clearer distinction between the rule and theory itself. Thus, the term norm seems like the most appropriate term to describe the selection rules for theories of causation. With these clarifications in mind, we are ready to examine how two groups of philosophers (reductivists and interventionists) rank theories of causation.

[2] These norms allow not just for binary but also for graded judgments: we can say that one theory is better than another theory as we will see in the next section.

[3] For example, some social scientists use quantitative methods which reveal their positivist commitments (Gattone, 2020). Others use qualitative methods which reveal their interpretivist commitments.

3 Reductivists

Reductivists are interested in theories of causation that provide an ontological reduction of facts about causation to some more fundamental facts (Paul and Hall, 2013, Chapter 2). This idea is evident in some of the most influential accounts of causation such as Lewis's counterfactual theory (1973). While these theories might seem very different at first, most reductivists share several common commitments (Paul and Hall, 2013, Chapter 2). To see why this is the case, we need to unpack reductivism. Consider the following biconditional theory[4] of causation:

C causes E if and only if $r(C, E)$ is true

Here C and E refer to particular events. For instance, let C denote the event when Gemma got her A.B. in economics from Harvard and let E denote the event when she became rich. In that case, the left-hand side of the biconditional reads: Gemma's A.B. from Harvard caused her to become rich. We will also assume that C temporally precedes En.

On the right-hand side, $r(C, E)$ denotes a dependence relation between events C and E. Here are two examples of such relations. It could be that $r(C, E)$ denotes the counterfactual 'had C not happened, E would not have happened' as in Lewis' original counterfactual theory (1973). If the counterfactual 'had Gemma not studied economics at Harvard, she would not have become rich' is true, we can conclude that Gemma's A.B. from Harvard caused her to be rich. On the other hand, it might be that events of type C are regularly followed by events of type E. Here $r(C, E)$ denotes an empirical regularity as in *some* versions of the regularity theories of causation (Paul and Hall, 2013, p.14). Perhaps most people who did an A.B. in economics at Harvard are rich, so an A.B. in economics from Harvard will also cause Gemma to become rich.

What these two examples have in common is that the expression $r(C, E)$ used on the right-hand side of the biconditional refers to something that is supposedly more

[4] Of course, a simple biconditional is unlikely to capture all the details of most theories of causation. For example, while we can write the counterfactual theory as a biconditional, we also need a semantic for evaluating counterfactuals. So, when I am expressing a theory of causation as a biconditional, it should be kept in mind that this is only a useful simplification.

fundamental than causation (e.g., counterfactuals). Why do reductivists insist on such an analysis instead of accepting that causation is primitive? The short answer is that most of them subscribe to a Humean ontology, in which we only observe separate events (Mumford and Anjum, 2011, pp.viii-ix). Whenever we *think* that event C causes event E, we observe only that event C is followed by event E. We do not actually observe if event C causes event E but only their temporal order. Since we do not observe actual causal relations, it must be the case that what we identify as causation are events C and E and their temporal order *plus* something else which is captured by $r(C, E)$ and could be regularities or counterfactuals. What matters is that $r(C, E)$ only contains things that are less fundamental than causation. For now, I will accept the Humean ontology but will revisit it in §6.a and §7.

While the assumption of a Humean ontology explains the motivation for seeking an ontological reduction, we might wonder how reductivists choose between two different theories of causation. In other words, what are the selection norms that guide their theory choice? Naturally, one of their norms states:

> *Reduction Norm.* A theory of causation is a good theory only if it is reductive.

But how can we select among two *reductive* theories? The preferred strategy is to use the 'method by counterexample' (Paul and Hall, 2013, pp.249-251). In this method, we evaluate if in various situations our theory of causation yields the same result as our intuitive causal judgment. If it does not, we will have evidence (albeit probably not decisive) against the theory. In that case, we need to either modify our theory of causation or argue why the counterexample is not problematic. We can roughly think of this process as analogical to hypothesis testing in science: we are testing our theory of causation (our hypothesis) on different tricky cases (our data) (Strevens, 2019, p.3).

For instance, consider Gemma's case:

> **Joint effects.** Gemma was born into an extremely wealthy family. She got into Harvard due to legacy admissions: many members of her family studied there. They helped her with her application and often made generous

> donations to Harvard. After getting her A.B. in economics, she decided to move to Bulgaria where she became a maths teacher in a middle school. While this was not a lucrative job, Gemma found working with children very fulfilling. A couple of years later she inherited 10 million dollars when her uncle died and so became rich. Overall, 95% of her wealth was inherited from her family.

Clearly, both Gemma's education and her wealth are the joint effects of her family's situation. So, our intuition tells us that Gemma's education is *not* the cause of her wealth. However, it is often the case that people who study for an A.B. in economics at Harvard end up rich. So, the regularity theory leads us to believe that Gemma must have also become rich because of her education. As we saw, however, this is not the case. Thus, we need to modify the simple version of the regularity theory in order to accommodate this counterexample.[5]

Gemma's case is not a problem for the counterfactual theory. The counterfactual 'had Gemma not studied economics at Harvard, she would not have become rich' is false. She would have still become rich due to family wealth. So, her A.B. from Harvard did not cause her to get rich. On the other hand, the counterfactual 'had Gemma not been born into a rich family, she would not have become rich' is true.[6] The counterfactual theory correctly judges that Gemma's wealthy family made her rich.[7]

The last two paragraphs showed how we can use the method by counterexample to select between theories of causation. Using this insight, we can formulate the norm:

> *Counterexamples Norm.* The fewer counterexamples our theory of causation faces, the better it is.

Note that we need to constrain the application of this norm to the set of cases, in which we can pronounce a clear causal judgment.[8] If there are cases, in which we cannot agree on a

[5] While this was a counterexample against the most simple regularity theory, the theory has more complex versions that can accommodate it (Paul and Hall, 2013, Ch. 1: §2.1, §3.3.1).

[6] Assuming she does not win the lottery, she embarks on a similar career path, etc. Due to space constraints, I am not discussing what is the right semantics for evaluating counterfactuals here.

[7] There are other counterexamples against the counterfactual theory of causation, e.g., late preemption cases (Paul and Hall, 2013, pp.99-143).

[8] We might constraint this set further to the cases discussed in Paul and Hall (2013) which include

judgment, we will also be unable to determine if our theory of causation yields the same prediction. Thus, we can conclude that when selecting between theories of causation, our choice should be driven by (i) the *Reduction Norm* and among the theories that are reductive by (ii) the *Counterexample Norm.*[9] While the *Reduction Norm* provides a necessary condition for a theory of causation, the *Counterexample Norm* allows us to rank how good different reductive theories are.

Should we accept reductivists' norms? In my view, these norms are not wrong *per se* but they are not the most fundamental norms. Given their norms, it is fairly clear why Paul and Hall should prefer certain theories over others. This is not the case for many other philosophers such as Woodward as we will see in §5. Nevertheless, both of Paul and Hall's norms are grounded in a more fundamental norm that I call the *Revised Functional Value Norm* and discuss in §§5-6. For the present moment, it suffices to say that reductivists incorrectly assume that their norms are the most fundamental norms. Once we introduce the *Revised Functional Norm*, we will see that they are just one set of norms among many others that should be acceptable to philosophers of causation.[10]

4 Interventionists

Another set of selection norms is offered by a group of philosophers that I call the *interventionists.* This group is sympathetic to Woodward's interventionist theory of causation (2003) or, at least, his general approach to causation (2014*a*; 2014*b*; 2015). In this section, we will first introduce Woodward's theory before examining another selection norm.

The interventionist theory is captured by the biconditional (Woodward, 2016):

Interventionist Theory. C causes E if and only if intervening on C in the

the most important examples from the literature. These cases include (e.g.) late preemption (Paul and Hall, 2013, Ch. 3) and causation involving omissions (2013, Ch. 4) among others.

[9] Paul and Hall (2013) identify several other conditions (and so candidates for selection norms) on what counts as a good theory. For instance, they reject theories of causation that reduce causation to extravagant metaphysical objects. In my view, these conditions are much less important than the norms we considered.

[10] This objection is not incompatible with Paul and Hall's own conclusion (2013, p.249): "Barring a fundamental change in approach, the prospects of a relatively simple and elegant ... theory of causation ... are dim."

right way is associated with a change in E

What does it mean to intervene 'in the right way'? Woodward (2015, §4) defines it as changing C in such a way that our intervention does *not* affect any of the other causes of E, apart from those causes that are lying on the causal path from C to E.[11] This requirement ensures that E changes because of the change in C and not because of the change in another cause of E.

To understand the notion of an intervention better, let us consider the causal claim 'Getting an MBA from Harvard will make you rich'. Suppose we set ethical and practical considerations aside and decide to test this claim via a Randomized Control Trial (RCT) (cf Woodward, 2015, §§4-5). In our RCT, we find a group of participants, randomly enroll half of them into Harvard's MBA, and compare their earnings several years later. If the participants with an MBA end up richer than those without, we can be sure that the causal claim is valid. Randomizing our treatment among our participants is an example of an ideal intervention in Woodward's sense. This is because we are minimizing the chance that the observed difference in earnings is due to another cause of earnings.[12] In other words, what explains the difference in earnings between the two groups is the MBA and so our causal claim is valid.

This RCT also suggests why many philosophers consider the interventionist theory as circular (Woodward, 2003, pp.104-105). To understand what causation is, we need to know what an intervention 'in the right way' is. But the notion of intervention requires us to know what the other *causal* factors affecting C are. For instance, in the RCT above we needed to know what the causes of earnings are when we randomize. So, we need to already have some notion of causation when we define an intervention. This is, however, circular: we defined causation in terms of intervention but defining intervention 'in the right way' requires us to know what causation is.

This circularity is problematic if we are looking for a reductive theory of causation expressed in the form of biconditional above. Suppose we express the interventionist

[11] Woodward (2003, p.98) lists the conditions for a legitimate intervention more formally.

[12] Note that I am not saying that the RCT produces two groups balanced in terms of all confounders except the treatment. While this 'balance assumption' often fails in practice, we can still say that randomization prevents 'systematic imbalances' in confounders among the two groups (Fuller, 2019).

theory via such a biconditional where $r(C, E)$ denotes 'intervening on C in the right way is associated with a change in E'. The trouble is that $r(C, E)$ here makes reference to causal notions in the definition of an intervention, so the interventionist theory does not provide a true reduction of causation (see also Strevens, 2007, 2008).[13] In that sense, the theory violates the *Reduction Norm*. Unsurprisingly, many reductivists reject it for this precise reason (Paul and Hall, 2013, p.26)

In response, Woodward (2014*a*,*b*, 2015) downplays the importance of circularity by arguing that the interventionist theory remains our most practical (or useful) theory. Clearly, then, reductivists and interventionists disagree about what makes for a good theory of causation. For interventionists such as Woodward, usefulness is a key feature of good theories which is not necessarily the case for reductivists. This idea is captured by Woodward's functional approach. We will now introduce this approach so that we can better understand how Woodward answers the circularity objection.

Woodward (2014*a*, pp.693-694) defines the functional approach as follows:

> [B]y a functional approach to causation, I have in mind an approach that takes as its point of departure the idea that [theories of causation] are sometimes useful in the sense of serving various [objectives] that we have. It then proceeds by trying to understand and evaluate various [such theories] in terms of how well they contribute to the achievement of these [objectives].

This passage says that we should strive to develop *useful* theories of causation. If we understand usefulness in terms of functional value, we can rank theories of causation on the basis of their functional value. This idea is captured in the following norm:

[13] We might worry that reductivists usually talk about single-case causal claims such as '*an MBA* will make *Arina* rich' whereas interventionists discuss generic causal claims such as '*MBAs* have made *many people* rich'. While it is true that interventionists usually discuss generic causation, their framework can also be utilized to analyze single-case causation (cf. Halpern and Pearl, 2005). Instead of the generic causal claim about MBAs, we could have used the single-case claim 'an MBA will make Arina rich' to illustrate an intervention 'in the right way'. Such an intervention would have ensured that no other causal paths to Arina's earnings is activated, e.g., via finding a rich partner or winning the lottery. I chose to use a generic causal claim for clarity, as the RCT example is more intuitive. However, I think that the different types of causal claims are not a problem for our comparison. Similar considerations apply to the worry that while the causal relata for interventionists are usually variables, the causal relata for reductivists are usually events.

> *Functional Value Norm.* The greater the functional value of a theory, the better the theory is.

This norm suggests that if a theory has greater functional value than another theory, the first theory ought to be selected. While Woodward refers to the norm as the functional approach (2014*b*), I call it the *Functional Value Norm* because it has the same purpose as reductivists' norms: to help us select between different theories of causation.

At this point, we might wonder how Woodward defines functional value (2014*a*). He defines it as the degree of usefulness of an object for achieving a particular objective. The best way to understand this definition is via an example. Your key is useful for unlocking the front door to your flat but it is (hopefully) not useful for unlocking the front door to my flat. So, your key has great functional value with respect to the objective of unlocking your door but no functional value with respect to another objective, namely unlocking my door. We can use the same idea to evaluate our theories of causation. We have various objectives related to our theories of causation such as explaining mental causation and distinguishing causation from correlation (Woodward, 2014*a*, §3). Depending on how well a theory achieves these objectives, it will have high or low functional value.

Woodward thinks that the interventionist theory is better at achieving these objectives than all other theories of causation. For instance, he argues that in contrast to most other theories of causation, the interventionist theory can explain why various statistical methods used for causal inference such as instrumental variables work (Woodward, 2015, §8). Thus, it has greater functional value and according to the *Functional Value Norm* it should be preferred. This reasoning provides Woodward's response to the circularity objection. While his theory might not be reductive, it has greater functional value than other theories and is thus more useful. So, if we accept the *Functional Value Norm*, we ought to prefer the interventionist theory.

5 Formulating the Revised Functional Value Norm

So far, we have seen two sets of selection norms: reductivists' *Reduction Norm* and *Counterexamples Norm* and interventionists' *Functional Value Norm.* While I consider the latter as the most important selection norm, in this section I show why it is *underspecified*, why this underspecification is a problem, and how we need to revise the norm to avoid this problem. I provide a positive argument in its favor in §6.

For interventionists, the *Functional Value Norm* is sufficient for selecting between two different theories. But what if we cannot agree upon a set of objectives that we want our theory of causation to achieve? In that case, it will be impossible to compare two theories on the same metric of functional value. The best way to see the power of this idea is to consider the following case:

> **Hypothetical disagreement.** Two philosophers Laurie and Jim have their own theories of causation. They are also interested in two different sets of objectives that are sufficiently different from each other. Both Laurie and Jim have chosen their own theory because it is the best at achieving the objectives that interest them. Jim then uses this fact to argue that his theory has the highest functional value and should be preferred over Laurie's theory.

Will Laurie be convinced by Jim's argument? I do not think so. Laurie can immediately respond that even though Jim's theory might be better at achieving the objectives he is interested in, Laurie is interested in different objectives. Given the objectives she cares about, her theory has greater functional value and thus should be preferred. In other words, Laurie does not accept Jim's estimates of the functional value of their theories and so does not accept his argument. Unless Jim can convince Laurie that his objectives are more important than her objectives, Laurie will not change her mind.

We can now see the underspecification of the *Functional Value Norm.* The norm is underspecified because it does not provide a list of objectives relative to which we can calculate functional value. Unless this is done, our estimates of functional value will not be comparable and we will not be able to use the norm for selecting between theories.

Even if Woodward's theory is the best at achieving the objectives he is interested in, this does not mean that it has the greatest functional value overall which is what his argument suggests. Thus, using the *Functional Value Norm* requires two things. First, it requires agreement on the fact that functional value is the only metric, on which we compare our theories of causation.[14] Second, it requires agreement on the set of objectives relative to which functional value is calculated.

How can we satisfy the second requirement? While there might be other options, our best bet is what I call the contextual strategy.[15] This strategy recognizes that there are many different communities thinking about causation and most such communities expect different things from their theory. The reductivists expect their best theories of causation to handle a lot of tricky counterexamples. In contrast, interventionists expect their theories to explain how causal inference in science works. According to the contextual strategy, we should recognize that different communities have different objectives and start thinking about functional value from this perspective. So, the contextual strategy resolves the underspecification by linking functional value to the objectives of a given community. In that sense, the functional value of a theory depends on the *context*, in which it is used.

We can now use the contextual strategy to formulate the norm:

> *Revised Functional Value Norm.* Given the objectives of a community, the greater the functional value of a theory, the better the theory is.

This norm implies that if a theory has greater functional value than another theory given the objectives of some community, the first theory ought to be selected by that community. When I refer to a community in this norm, I mean a group of thinkers about

[14] Does this requirement suggest a pragmatic theory of truth? To some extent, yes. I discuss this question further in §7.

[15] One alternative is the *ideal theory strategy.* It resolves the underspecification by linking the calculation of functional value with the objectives of the ideal theory that would mark the end of all philosophical speculation on causation. While we do not have access to this theory, identifying its characteristics will allow us to get a set of objectives, relative to which we can calculate functional value. For example, the ideal theory will not suffer from late preemption cases. The trouble with this strategy is that it requires us to agree on some of the fundamental characteristics of the theory, e.g., if it is reductive. Given how much we disagree about causation, this seems unlikely. So, the ideal theory strategy is unlikely to deliver a set of objectives that all of us can accept.

causation who do not have to be philosophers (e.g., they could be methodologists) and who do not have to self-identify as belonging to such a community. Instead, what defines a community for our purposes is the presence of *enough* shared objectives. As long as our collection of thinkers is united by a bare minimum of such objectives, we can always list some of these objectives and calculate functional value relative to them. Note that our definition of a community also does not prevent us from thinking about the objectives of a single-thinker community (e.g., just Woodward's objectives), even though the word community entails a *group* of thinkers.[16]

6 Motivating the revised functional value norm

Why should we accept the *Revised Functional Value Norm*? In this section, I offer two positive arguments in its favor. First, this norm grounds the norms used by other communities. Second, it explains not just why philosophers disagree about causation but also why methodologists disagree with philosophers about causation.

6.a Reformulate reductivists' norms

As we saw, reductivists select between theories using the *Reduction Norm* and the *Counterexamples Norm*. In particular, reductivists are interested in developing a theory of causation that (i) reduces causation to something more fundamental and (ii) accommodates many counterexamples. But notice that (i) and (ii) are also the main objectives of reductivists. This insight matters because it suggests that we can calculate the functional value of different theories given reductivists' objectives.

Consider then the following two norms:

(1) Any *reductive* theory has greater functional value than all *non-reductive*

[16] There are clear similarities between the contextual strategy and *causal perspectivism* (Price, 2005). For instance, both approaches recognize that two different communities might end up with two completely different theories of causation. However, notice that Price is not interested in theories of causation but in causal claims. While I provide an analysis of why we accept different theories of causation, Price focuses on why we make different causal claims. As a result, Price is committed to denying the objectivity of causal claims (cf Ismael, 2016). As I argue below (§6), my account does not commit me to this position, even though it *might* commit me to a denial of the objectivity of *theories of causation*.

theories.

(2) A reductive theory has greater functional value than another reductive theory if it faces fewer counterexamples.

Norm (1) restates the *Reduction Norm* in terms of functional value whereas norm (2) restates the *Counterexamples Norm* in terms of functional value. For reductivists, it should not matter whether they use (1) and (2) or the original norms because (1) and (2) are based on the same two objectives. Crucially, norms (1) and (2) show us why the *Revised Functional Value Norm* is the most fundamental norm.

Let us express this idea more precisely. What I mean by the last claim is that the *Revised Functional Value Norm* ground all other selection norms for theories of causation. We can always use facts about the objectives of a community and this norm to derive all of its other norms. This relation is asymmetric: from the norms used by that community, we cannot derive the *Revised Functional Value Norm* and their objectives. Given that grounding is a huge topic in contemporary metaphysics (Bliss and Trogdon, 2021), this argument requires a lot more work.[17] Unfortunately, I cannot do it full justice here due to space constraints. However, the connection between (relative) fundamentality and grounding helps us to understand why the *Revised Functional Value Norm* is so important. This idea also shows why Paul and Hall's approach is not entirely satisfactory as argued in §3. While the reductivists do a great job at clarifying the norms they use to rank theories of causation, it remains unclear not only why their norms are the only norms we can use to discover the nature of causation but also what the most fundamental norm is, i.e., the *Revised Functional Value Norm.*

One way to object to my argument here is to ask if norms (1) and (2) truly capture the meaning of reductivists' selection norms. A reductivist might concede that accommodating counterexamples is an objective of our theory. But is offering a reductive theory really *an objective*? Instead, we might think that it is a theoretical virtue of our theory and not really an objective that a theory can achieve. A theory is by assumption either

[17] For example, it concerns conceptual rather than metaphysical grounding (Bliss and Trogdon, 2021, §1.3.) and understands grounding in terms of explanation rather than determination (Bliss and Trogdon, 2021, §1.1.).

reductive or not and it does not make sense to talk about achieving the objective of becoming a reductive theory. In my view, however, the *Reductive Norm* does point towards an objective: this is the objective of providing a theory of causation consistent with a community's ontology. If (like reductivists) we have a Humean ontology, in which causation is not primitive, we will naturally prefer a reductive theory. If we have an ontology, in which certain causal links are primitive, reductive theories are not necessarily more valuable than non-reductive ones.

While Woodward does not go as far as embracing primitivism, some of his recent work does suggest that we cannot hope to understand causation whilst avoiding causal claims (Weinberger et al., 2023). So, he seems to be an anti-reductionist about causation. Given this ontological commitment, he is not after reductive theories. This suggests that reformulating the *Reduction Norm* using the *Revised Functional Value Norm* and the objective of providing a theory consistent with a community's ontology makes sense. There is no reason to think that our norm cannot ground the *Reduction Norm*.

6.b Explaining non-philosophical disagreements

We saw above that different philosophers subscribe to different theories of causation because they subscribe to different norms. Since we can use the *Revised Functional Value Norm* to reformulate each community's norms in terms of functional value, the disagreements boil down to a disagreement about the objectives of different communities. In other words, interventionists and reductivists rank theories of causation in different ways because they have different objectives. We can use the same idea to explain why scientists disagree with philosophers about causation. This provides another argument in favor of the *Revised Functional Value Norm*.

To see why, note that reductivists are not the only community interested in counterfactual theories of causation. A lot of methodologists[18] are fond of the so-called *potential outcomes framework* (Imbens and Rubin, 2015) which can be interpreted as a coun-

[18] Methodologists in this paper refer only to those methodologists who use this framework and not those who use Pearl's Causal Model. There is an ongoing debate between advocates of the two approaches (Weinberger, 2022).

terfactual theory. This framework begins by defining two potential outcomes for each individual: the first one gives us the value of the outcome of interest *with* treatment and the second gives us the value of the outcome of interest *without* treatment. For example, the first potential outcome can be my earnings *with* an MBA from Harvard and the second potential outcome can be my earnings *without* this MBA. The difference in the two potential outcomes gives the causal effect of the MBA on my earnings.

Crucially, we can interpret the two potential outcomes as counterfactuals. The first counterfactual tells us how much I would have earned, had I done an MBA, whereas the second counterfactual tells us how much I would have earned, had I not done an MBA. The causal effect is then the difference in earnings in the two counterfactual worlds. If the counterfactual 'I would have earned much less, had I not done an MBA' is true, then we can be sure that the MBA will cause an increase in my earnings. So, the *potential outcomes framework* can be read as a counterfactual theory and belongs to the same family of theories as reductivists' counterfactual theory.

Why is this insight important? Because the two counterfactual theories look nothing alike. Methodologists use their version to develop new estimators for causal inference and study the properties of these estimators formally. For this reason, the potential outcomes framework is well integrated with statistical theory (Imbens and Rubin, 2015). On the other hand, reductivists seek a semantic for evaluating the truth conditions of counterfactuals, given that (classic) propositional logic cannot be used for this purpose (Starr, 2022, §1). They also spend a lot of time exploring if the counterfactual theory can accommodate tricky counterexamples via neuron diagrams. Such diagrams or graphs are a tool that some methodologists do not find helpful (Pearl, 2009; Imbens, 2020). Thus, we can see that the two communities have developed very different theories of causation, despite starting from similar starting points.

We can make sense of these differences via the *Revised Functional Value Norm*. Reductivists and methodologists have different objectives. Most reductivists are probably not particularly interested in the asymptotic properties of the two-stage least square estimator whereas most methodologists are unaware of the problems raised by late preemption. Relative to the latter objective, the potential outcomes framework has very little functional value but relative to the former objective, it has very high functional value. As a result, it is valued highly by methodologists but not so much by reductivists. It is thus not a surprise that the two communities end up with very different theories of causation, even though both started with a counterfactual understanding of causation. They simply had different objectives which led them to value different features of a theory of causation. In that sense, the *Revised Functional Value Norm* helps us also explain non-philosophical disagreements about causation.

7 Objections

Now that we have introduced the *Revised Functional Value Norm*, let me reply to three worries we might have about it: i) it allows for communities with peculiar objectives and theories of causation, ii) it denies the objectivity of causal claims and iii) it commits us to a form of relativism about ontological reductions of metaphysical concepts.

7.a Communities with peculiar objectives

We might worry that the *Revised Functional Value Norm* does not allow us to make judgments on communities that adopt a theory of causation for peculiar reasons. To see why this is an issue, consider an imaginary cult of Malebranche.[19] Its members believe that Malebranche was the greatest philosopher that ever lived and their sole objective is to convince the world of this fact. Unsurprisingly, the cult accepts Malebranche's occasionalism[20] as the best theory of causation. After all, causation is a deep philosophical problem and the greatest philosopher that ever lived will surely have the right answer.

If we accept the *Revised Functional Value Norm*, we have to admit that there is nothing wrong with this reasoning. Given the objective of showing that Malebranche is

[19] Malebranche was chosen for no particular reason. I could have used any other thinker who offered an original theory of causation, e.g., Patrick Suppes.

[20] Occasionalism is the doctrine that only God has real causal powers and all other causes are not real Lee (2020). We can formulate occasionalism as a biconditional theory of causation (as discussed in §3) where $r(C, E)$ refers to divine volition. However, it cannot be a reductive theory of causation unless we stop assuming a Humean ontology.

the greatest philosopher, occasionalism has the greatest functional value. A reductivist can easily deny that the cult's reasoning makes sense. Since occasionalism violates the *Reduction Norm*, it is not a good theory. This seems to be an advantage of reductivism: it seems unreasonable to unquestioningly embrace my favorite philosopher's theory of causation just because I want to convince you how great she is. So, in contrast to the *Revised Functional Value Norm*, reductivists' norms allow us to judge communities that adopt a theory of causation for peculiar reasons.

While it is true that our selection norm says nothing about what is best *for* the cult, we can still use it to say that *for* most other communities thinking about causation Malebranche's occasionalism is not very useful. After all, most communities do not share the objectives of Malebranche's cult. In addition, occasionalism cannot explain why causal inference techniques work and does not reduce causation to something which is more fundamental in a Humean ontology. We can still use our norm to say that occasionalism has no functional value for reductivists and for interventionists. So, it is not such a big problem that our norm prevents us from rejecting the cult's reasoning.

7.b Objectivity of causal claims

A skeptic about the *Revised Functional Value Norm* might raise one more objection: the norm denies the objectivity of causal claims. Causal claims are objective if their truth can be verified regardless of the presence of specific agents and their beliefs. Denying the objectivity of causal claims is problematic because it undermines various claims in science that are taken to hold independently of who the observer is or if there are any observers at all. Objects on the dark side of the Moon, for instance, are still affected by its gravity, even when we do not observe it.

Why would the *Revised Functional Value Norm* be denying the objectivity of causation? We can see this most clearly via an example.[21] Suppose that some scientific community uses its best theory of causation T_1 to make a set of causal claims about their domain of interest. These causal claims are considered to be approximately true because

[21] Admittedly, my discussion hinges on this specific example. However, the main point should still stand: the *Revised Functional Value Norm* does not have to undermine the objectivity of causation.

they are made in line with the best theory T_1. Since the community follows the *Revised Functional Value Norm*, it has selected T_1 because it achieves a lot of their objectives. Let us suppose further that if these objectives had been slightly different, another theory T_2 would have been chosen. Crucially, the set of causal claims that are considered valid by T_2 is *not* identical to the set of causal claims that are considered valid by T_1. In reality, the community selected the initial set of objectives that led them to choose T_1. However, in another close possible world, it could have easily selected T_2 and ended up with a different set of valid causal claims.

This example seems to suggest that the set of causal claims considered valid in our community does not depend on objective facts about the world but on the objectives that were initially chosen by the community. Unfortunately, it is easy to imagine how the values of individual members of the community can influence the choice of objectives. For instance, a very charismatic scientist might have exerted a disproportionate amount of influence on that choice. The values of scientists affect which theory is chosen and thus which causal claims are considered as true. In that sense, the validity of causal claims in the community depends on non-objective factors. The skeptic can interpret this result as evidence that the *Revised Functional Value Norm* denies the objectivity of causal claims. If the charismatic scientist was not there, a different set of objectives would have been selected, another theory selected and different causal claims considered valid.

One way to reply to this objection is to say that our norm does not aim to tell us things about the metaphysics of causation directly. It allows both communities that deny and accept the objectivity of causation to offer their own theories. It allows for this because it tries to make no strong assumptions about causation. Its main purpose is to evaluate existing theories, not to make any direct contributions to the metaphysics of causation. Any contributions to the latter should come as a byproduct of using the norm in a particular community. For this reason, accusing the contextual strategy of denying the objectivity of causation misrepresents its purpose.

Even if our skeptic does not accept this reply, *Revised Functional Value Norm* does not necessarily deny the objectivity of causation. Consider once again the example the skeptic used to introduce her objection. Assume that the validity of causal claims in the domain in question is objectively determined by the theory of causation T^*. Unfortunately, T^* is not available to us. Perhaps, T^* reduces causation to something ontologically fundamental which we have not discovered yet. So, we need to pick one of our currently available theories to make causal claims in that domain. This would be the theory that approximates T^* most closely out of all existing theories. Depending on the set of objectives we pick, the *Revised Functional Value Norm* might recommend either T_1 or T_2. However, this does not undermine the fact that the validity of causal claims in the domain in question is objectively determined. It is just that different theories of causation approximate T^* in different ways and so give different predictions about the validity of certain causal claims.

If this sounds a bit too abstract, suppose there exists a domain, in which there are no preemption cases. The counterfactual theory is T^* in this domain. This means that it correctly determines the set of correct causal claims. Unfortunately, we are studying the domain in question in the 19th century. At that time philosophers have not yet discovered the counterfactual theory. So, the existing theories of causation are unable to identify the correct set of causal claims but that does not mean that such a set does not exist. For this reason, I think that the *Revised Functional Value Norm* by itself does not have to undermine the objectivity of causation. More broadly, this argument suggests that our norm is compatible with a range of views about the status of causal claims in the physical world, including Strevens' fundamental causal network (2013) and Norton's causal skepticism (2003).[22]

7.c Other metaphysical disagreements

We might worry that the *Revised Functional Value Norm* challenges our ability to resolve a lot of disagreements in metaphysics. The arguments in §6.b suggested that philosophers disagree about causation because they want different things from their theories. Does

[22] The main reason is that our norm operates on the level of causal talk and does not have to make any claims about causal process in the world. So, the validity of the norm does not depend on whether causation is fundamental in the physical world (like Strevens thinks) or not (like Norton thinks).

the same logic extend to other disagreements in metaphysics? After all, philosophers disagree fiercely about most concepts in metaphysics such as free will, grounding, the self, God, etc. Take free will as an example. Incompatibilists think that free will and determinism are mutually exclusive which is denied by compatibilists. If we analyze this debate in a similar vein to the debate between interventionists and reductivists, we might be committed to saying that compatibilists and incompatibilists disagree because they want different things from free will. An ontological reduction[23] of free will that will satisfy both compatibilists and incompatibilists can thus never be performed. So, we should just give up debating free will. Intuitively, this seems like an unsatisfying implication of my argument, as it challenges our ability to resolve most metaphysical debates.

One response to this worry is to say that causation is fundamentally different from other metaphysical concepts and thus less malleable to an ontological reduction. Why would that be? While some theories of causation (e.g., Lewis' counterfactual theory) are largely metaphysical, others (e.g., the potential outcomes framework) are largely epistemological. So, when we try to do an ontological reduction of causation, we might conflate metaphysical and epistemological aspects of it. In contrast, a lot of traditional metaphysical concepts such as free will remain largely metaphysical. There are, of course, epistemological questions about free will. For example, does quantum mechanics show that free will exists? Similarly, does it show that *causation* does not exist? These are both legitimate epistemological questions but they remain questions about the *nature* of causation and free will.

In contrast, the epistemological questions I have in mind concern the fact that free will has not been formally operationalized in the same way as causation. There is no equivalent to causal modeling or the potential outcomes framework in the free will literature. We can use causal modeling to make sense of metaphysical explanation and causal inference in social science. This gives rise to a distinctive set of epistemological questions such as how essential the faithfulness assumption is for causal modeling. Such questions are

[23] Following (Paul and Hall, 2013, pp.25-38), I am convinced that we should strive for an ontological reduction of causation. However, I could have also used the term conceptual analysis here and the discussion would not have changed.

absent in the debates about free will because free will has not been operationalized in the same way. Thus, aiming to provide an ultimate ontological reduction makes more sense in the case of free will but not in the case of causation.

We might worry that this response simply begs the question. Suppose we limit ourselves to *metaphysical* questions about causation. Does my argument imply that an ontological reduction of causation in this case will also not work? This implication does not necessarily follow. My argument does allow for an ontological reduction of causal claims within particular communities. It is perfectly fine for Paul and Hall to disagree with other reductivists about causation. More generally, so long as a community shares enough objectives, it can perform a meaningful ontological reduction. This might even be the best way to sort out their disagreements, given that the difference in their objectives is not too big. At any rate, we should acknowledge that an ontological reduction might apply only in a specific context, and generalizing its results to other domains will not always be possible.

In fact, there are metaphysical debates that accept this perspective. While both sides claim to talk about concept x, they acknowledge that they mean different things by x and concede that meaningful ontological reductions are possible in a more limited sense. Consider debates about the self (Schechtman, 2007, pp.1-3). The main objective of personal identity theorists is to explain the persistence of the self through time. In contrast, the main objective of narrative theorists is to identify the set of beliefs, values, and desires that define my self. While the two literatures share certain similarities, they seem to be pursuing very different theories of the self. Thus, it is fine if narrative theorists perform an ontological reduction of their own concept separately from personal identity theorists, and *vice versa*. Of course, it goes without saying that the two sides can draw on each other to improve their own theories but neither should claim to solve all problems about the self and provide the ultimate theory of the self.

8 Conclusion

We started by asking why do philosophers disagree about what makes a good theory of causation. I suggested that different philosophers subscribe to different selection norms when deciding between two theories of causation. Thus, reducitivists and interventionists disagree about causation because they have different selection norms. One example of such a norm is Woodward's *Functional Value Norm*. While this norm suffers from underspecification and should be rejected, I offered a modified version that can deal with this problem: the *Revised Functional Value Norm*. This norm states that given the objectives of a community, the greater the functional value of a theory, the better the theory is. I then argued that the *Revised Functional Value Norm* is the most fundamental selection norm, as it grounds the norms of other communities. Using this norm we can also explain why not only philosophers but also methodologists disagree about causation. Their disagreements boil down to the fact that they have different objectives. We also saw that accepting the *Revised Functional Value Norm* does not require us to either deny the objectivity of causal claims or our ability to resolve other metaphysical disagreements.

References

Bliss, R. and Trogdon, K. (2021), Metaphysical Grounding, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2021 edn, Metaphysics Research Lab, Stanford University.

Fuller, J. (2019), 'The confounding question of confounding causes in randomized trials', *The British Journal for the Philosophy of Science* .

Gattone, C. F. (2020), *A Balanced Epistemological Orientation for the Social Sciences*, Lexington Books.

Halpern, J. Y. and Pearl, J. (2005), 'Causes and explanations: A structural-model approach. part i: Causes', *The British journal for the philosophy of science* .

Henne, P. (2023), Experimental metaphysics: Causation, *in* A. M. Bauer and S. Kornmesser, eds, 'The Compact Compendium of Experimental Philosophy', De Gruyter, pp. 133–162.

Hitchcock, C. and Knobe, J. (2009), 'Cause and norm', *Journal of Philosophy* **106**(11), 587–612.

Horwich, P. (1991), 'On the nature and norms of theoretical commitment', *Philosophy of Science* **58**(1), 1–14.

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), 1129–1179.

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

Ismael, J. (2016), 'How do causes depend on us? the many faces of perspectivalism', *Synthese* **193**, 245–267.

Knobe, J. and Fraser, B. (2008), Causal judgment and moral judgment: Two experiments, *in* W. Sinnott-Armstrong, ed., 'Moral Psychology', MIT Press.

Lee, S. (2020), Occasionalism, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2020 edn, Metaphysics Research Lab, Stanford University.

Lewis, D. (1973), 'Causation', *The journal of philosophy* **70**(17), 556–567.

Littlejohn, C. (2013), The dual-aspect norms of belief and assertion: A virtue approach to epistemic norms, *in* C. Littlejohn and J. Turri, eds, 'Introduction', Oxford Uni Press.

Mumford, S. and Anjum, R. L. (2011), *Getting causes from powers*, Oxford University Press.

Norton, J. (2003), 'Causation as folk science', *Philosophers' Imprint* **3**, 1–22.

Paul, L. and Hall, N. (2013), *Causation: A user's guide*, Oxford University Press.

Pearl, J. (2009), 'Myth, confusion, and science in causal analysis'.
URL: *https://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf*

Price, H. (2005), Causal perspectivalism, *in* H. Price and R. Corry, eds, 'Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited', Oxford University Press.

Pritchard, D. (2013), Epistemic luck, safety, and assertion, *in* C. Littlejohn and J. Turri, eds, 'Epistemic Norms: New Essays on Action, Belief, and Assertion', Oxford University Press.

Schechtman, M. (2007), *The constitution of selves*, Cornell University Press.

Starr, W. (2022), Counterfactuals, *in* E. N. Zalta and U. Nodelman, eds, 'The Stanford Encyclopedia of Philosophy', Winter 2022 edn, Metaphysics Research Lab, Stanford University.

Strevens, M. (2007), 'Review of Woodward, "Making Things Happen"', *Philosophy and Phenomenological Research* **74**(1), 233–249.

Strevens, M. (2008), ‘Comments on Woodward’s ‘Making things happen’’, *Philosophy and Phenomenological Research* **77**(1), 171–192.

Strevens, M. (2013), ‘Causality reunified’, *Erkenntnis* **78**(2), 299–320.

Strevens, M. (2019), *Thinking off your feet: How empirical psychology vindicates armchair philosophy*, Harvard University Press.

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

Weinberger, N., Williams, P. and Woodward, J. (2023), ‘The worldly infrastructure of causation’, *Phil-Sci Archive* .
URL: *http://philsci-archive.pitt.edu/22125/*

Williamson, T. (2000), *Knowledge and its Limits*, New York: Oxford University Press.

Woodward, J. (2003), *Making things happen: A theory of causal explanation*, Oxford university press.

Woodward, J. (2014*a*), ‘A functional account of causation’, *Philosophy of Science* **81**(5), 691–713.

Woodward, J. (2014*b*), Interventionism and the missing metaphysics: A dialog, *in* M. Slater and Z. Yudell, eds, ‘Metaphysics and the Philosophy of Science: New Essays’, Oxford University Press, pp. 193–228.
URL: *https://philpapers.org/rec/WOOIAT-8*

Woodward, J. (2015), ‘Methodology, ontology, and interventionism’, *Synthese* **192**(11), 3577–3599.

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