Interventionism for the Intentional Stance: True Believers and Their Brains

Markus I. Eronen

m.i.eronen@rug.nl

Abstract

The relationship between psychological states and the brain remains an unresolved issue in philosophy of psychology. One appealing solution that has been influential both in science and in philosophy is the intentional stance developed by Daniel Dennett, according to which beliefs and desires are real and objective phenomena, but not necessarily states of the brain. A fundamental shortcoming of this approach is that it does not seem to leave any causal role for beliefs and desires in influencing behavior. In this paper, I show that intentional states ascribed from the intentional stance should be seen as real (interventionist) causes, develop this to an independently plausible ontological position, and present a response to the latest interventionist causal exclusion worries.

1. Introduction

One of the key issues in the philosophy of psychology is the relationship between intentional states, such as beliefs, and neurobiological states of the brain. Daniel Dennett's (1971, 1987, 1991, 1996, 2009) theory of the intentional stance (also known as the intentional strategy or intentional systems theory) is an attractive account of this relationship that has provoked wide-ranging debates in philosophy, and has been influential in science as well (see, e.g., Gergely et al. 1995; Gallagher et al. 2002; Griffin and Baron-Cohen 2002). In a nutshell, the intentional stance approach consists of treating the object whose behavior one wants to predict as a rational agent with beliefs and desires and other intentional states (Dennett 1987, 15).[1] When this strategy is successful, the agent is a "true believer" and really has beliefs and desires. This leads to a picture where beliefs and desires are real and objective phenomena, but not necessarily states of the brain, thus resulting in an appealing middle position between strong realism and instrumentalism.

However, the intentional stance has not been broadly accepted as an account of the nature of intentional states, most importantly because it appears to make them just abstract objects that have no causal powers to influence behavior, as opposed to the underlying brain states that do the causal work (Horgan and Woodward 1985; McCulloch 1990; Pöyhönen 2014; Rey 1994; Slors 2007; Zawidzki 2012). In contrast, I will argue in this paper that the recently popular interventionist account of causation provides a way of understanding how Dennettian intentional states can have genuine causal roles without necessarily being states of the brain. I will also develop this to an independently plausible ontological position that leads to fruitful connections to other issues in contemporary philosophy of science, and then address worries of interventionist causal exclusion.

It should be noted that the position developed here is not intended to capture all the features of Dennett's original theory. Dennett has argued, for example, that beliefs and desires should be seen as *abstracta* that are equivalent to behavior attributions, and that his theory is a sort of holistic logical behaviorism (Dennett 1987, ch. 3), which is clearly in conflict with what I defend. Due to this and other differences, the account in this paper should be seen as a new (more realistic) version of the intentional stance theory.

---

[1] The intentional stance can be seen as an account of many things (Zawidzki 2012): intentionality, psychological explanation, the nature of beliefs and desires, or everyday social cognition; my focus here will mainly be on the intentional stance as an account of the nature of beliefs and desires (and other intentional states) and their relation to the brain. The intentional stance is also often taken to represent the "theory-theory" approach to social cognition, as opposed to "simulation theory" or "embodied social cognition" approaches (Zawidzki 2012). I do not intend to enter this debate here. Assuming that beliefs and desires play an important role in psychological explanation (and I take a large proportion of philosophers of psychology to accept this assumption), we need an account of their causal and ontological status, and this paper provides one such account.

In the next section, I will go through the original intentional stance account in more detail. In section 3, I will briefly show that the interventionist theory of causation allows intentional states ascribed from the intentional stance to be real causes. In section 4, I will argue that this approach retains the most attractive ontological features of Dennett's intentional stance theory, and in section 5, I will defend my account against the interventionist causal exclusion argument. Section 6 consists of concluding remarks and suggestions for further research.

2. The Intentional Stance

Dennett's starting point is that there are various strategies to predict the behavior of an object or a system.[2] A powerful and common strategy is to take the *physical stance*: to find out the structure and constitution of an object and the forces acting upon it, and then use physical laws or regularities to predict its behavior. Sometimes a more efficient strategy is to take the *design stance* and to treat the object as having been designed for some purpose: For example, I know that my cell phone has been designed to make phone calls, so I can reliably predict that it if I press the right buttons it will call my partner, and I can do this without knowing how the phone is constituted and what are the physical forces acting upon it. Insofar as biological entities can be seen as having been designed by evolution, the design stance can also be applied to them.

Let us then consider the prediction of human behavior. Jim utters to Jane: "I am going to get a cup of coffee now". Predicting Jim's subsequent behavior from the physical or design stance would be extremely difficult or even practically impossible, but Jane can adopt the *intentional stance*, and treat the object (the other person, in this case Jim) as a rational agent with beliefs and desires (and other intentional states). Jim is rational, he desires coffee, and there is a pot of fresh coffee in the kitchen, so Jane predicts that he will go to the kitchen. When predicting the behavior of the agent from the intentional stance, one first figures out what beliefs and desires it ought to have given its situation and purpose, and then reasons what the agent ought to do in this situation to further its goals. In most cases, what the agent ought to do is also what she/he/it will do.

This intentional strategy seems to be widely used and remarkably powerful: 'Do people actually use this strategy? Yes, all the time … And when does it work? It works with people almost all the time' (Dennett 1987, 21). Arguably, the intentional strategy is also prominent in science: Explanations in social science and psychology are largely based on treating humans as rational agents and attributing them beliefs and desires, and even in many branches of biology (such as ethology), the behavior of

---

[2] The following overview is based on 'True Believers: The Intentional Strategy and Why It Works' (Dennett 1987, ch. 2), which is generally considered to be the standard version of the account.

animals is often predicted from the intentional stance (Dennett 1987, ch. 7; Dennett 2009). It is a matter of debate whether the intentional strategy really is as generally successful as Dennett claims, but as the aim in this paper is to show that Dennett's account can be made conceptually consistent, I will treat the empirical success of the intentional strategy as a background assumption (see also Dennett 1987, ch. 4 for his rebuttals).

Perhaps the most interesting aspects of the intentional stance are its ontological implications. It is supposed to avoid the realistic view that beliefs and desires are real things in the head, but also the sort of relativism or interpretationism where the question whether someone has a certain belief is merely a matter of interpretation or perspective. In Dennett's own words: 'My thesis will be that while belief is a perfectly objective phenomenon (that apparently makes me a realist), it can be discerned only from the point of view of one who adopts [the intentional] strategy, and its existence can be confirmed only by an assessment of the success of that strategy (that apparently makes me an interpretationist)' (Dennett 1987, 15). Moreover, "*all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there* is to really and truly believing that *p* (for any proposition *p*) is being an intentional system for which *p* occurs as a belief in the best (most predictive) interpretation" (ibid., 29).

Beliefs and desires are objective phenomena because there are *real patterns* underlying beliefs, desires and behavior (Dennett 1987, 39-40; Dennett 1991). However, these real patterns are not visible from the physical or design stance, so the only way to ascertain whether an object has beliefs or desires is to adopt the intentional stance and to see whether the behavior of the object can be reliably predicted from the intentional stance (ibid.).

The main objection to this account of the ontological nature of beliefs and desires is that it cannot escape instrumentalism, as it seems to imply that beliefs and desires are not the real causes of behavior (Horgan & Woodward 1985; McCulloch 1990; Pöyhönen 2014; Rey 1994; Slors 2007; Zawidzki 2012). In *The Intentional Stance*, Dennett embraces this conclusion (Dennett 1987, 71; see also Dennett 1987, 54-57). However, it is deeply problematic. If beliefs do not have causal powers, why should we think that they are real? They may appear in interpretations that are predictively useful, but from an ontological point of view, they do not seem to be doing anything at all, and consequently their presence or absence cannot as such have any influence on the behavior of an object. The view that beliefs and desires are not real causes also goes against the commonsensical and widely held view (going back at least to Davidson 1963) that reasons are causes of behavior. Thus, far from being a form of realism, Dennett's intentional strategy seems to amount to a form of instrumentalism or epiphenomenalism.

Probably for these reasons, in later publications Dennett has suggested that intentional states can in fact have causal roles. For example, in a footnote to the article 'Real Patterns', he writes: 'Several interpreters of a draft of this article have supposed that the conclusion I am urging here is that beliefs (or their contents) are epiphenomena having no causal powers, but this is a misinterpretation traceable to a simplistic notion of causation. If one finds a predictive pattern of the sort just described one has ipso facto discovered a causal power – a difference in the world that makes a subsequent difference testable by standard empirical methods of variable manipulation' (Dennett 1991, footnote 22). In a later response to critical comments (Dennett 2000, 357-358), he emphasizes that we need a concept of causation that can accommodate higher-level causes such as beliefs and centers of gravity. Unfortunately, Dennett does not connect this view to any theory of causation, or give a clear argument for treating intentional states as real causes, which leaves these brief remarks hanging in the air.

As I will now proceed to show, states ascribed from the intentional stance can be seen as interventionist causes.[3] As this part draws from existing work on applying interventionism to mental causation, I will go over it rather quickly, and then move on to discuss the ontological consequences of interpreting Dennettian intentional states as interventionist causes, and possible objections to this approach.

3. Interventionism and the Intentional Stance

The interventionist account (or simply interventionism) has been developed by James Woodward (2003), building on earlier work on causal modeling (Pearl 2000, Spirtes, Glymour and Scheines 1990), and is becoming increasingly popular in philosophy of science and elsewhere, but has not yet been applied to the intentional strategy. Its core idea is that causes make a difference for their effects, and causal relationships are relationships that are potentially or ideally exploitable for manipulation and control. More precisely, a necessary and sufficient condition for X to be a cause of Y (in a causal representation with a set of variables **V)** is that the value (or the probability distribution) of Y would change under some intervention on X, when all other variables in **V** (that are not on the path from X to Y) are held fixed (Woodward 2003). Interventions have to satisfy specific conditions: An intervention on X with respect to Y has to cause the change in X; the change in X has to be entirely due to the intervention and not any other factors; the intervention should not change Y directly; and it should be uncorrelated with any causes of Y that are not on the path from X to Y (see

---

[3] Slors (2007) also defends the view that Dennettian intentional states can be causes, but he is drawing from the notion of causal relevance as defined by Jackson & Pettit (1990) instead of interventionism, and his approach and conclusions are very different from mine.

Woodward 2003, ch. 3, for more details). In a nutshell, what these conditions imply is that the intervention has to change X in such a way that the change in Y is only due to the change in X and not any other factor (Woodward 2015b).[4]

These abstract definitions are easier to grasp with the help of an example. Let us suppose that we want to find out whether a drug (e.g., penicillin) causes recovery from a disease (e.g., staphylococcal infection).[5] We can represent administering the drug with variable A (value 1 = drug administered, value 0 = no drug administered) and recovery from the disease (e.g., within 2 weeks) with variable R (value 1 = recovery, value 0 = no recovery). In order to assess whether A causes R (in a population of infected individuals), we need to intervene on A while holding all other variables (that are not on the path from A to R) fixed and see if there is a change in (the probability of) R. Often this is done in practice by randomized controlled trials: Subjects are divided into two groups that are as similar as possible, the difference being that the one group receives the actual drug, while the other group receives a placebo. The intervention should satisfy the conditions outlined above, that is, it should change A in such a way that the change in R is only due to the change in A and not any other factor – for example, if the process of administering the drug involves other procedures that are beneficial for recovery, the intervention is not of right kind, and we cannot draw the conclusion that the drug causes recovery.

As interventionism is not intended to provide a conceptual analysis of causation (as in, for example, Lewis 1973), one may wonder whether it is suited for the philosophical task of evaluating the causal status of Dennettian intentional states. More generally, not everyone agrees that interventionism is a good or satisfactory account of causation at all, for example due to problems of infinite regress in defining causes and interventions (see, e.g., Baumgartner 2009a). However, interventionism can be seen as a functional or methodological account that should be evaluated by its usefulness, and it has already proven to be useful in illuminating the nature of causal relations (as they appear in science), and in giving criteria for distinguishing causal relations from other relations, most importantly correlations (Woodward 2014, 2015a, 2015b).[6] Interventionism also has the advantage of being methodologically relevant and reflecting scientific practices of causal modeling and reasoning (ibid.).

---

[4] Note that interventions do not necessarily involve human agency or manipulation; also 'natural' interventions can satisfy these conditions. In interventionism, the relata of causation need to be represented as variables, but this does not put any substantial metaphysical constraints on the relata – for example, properties can be represented as binary variables, so that value 1 represent the presence of the property and value 0 its absence.
[5] This is an adapted version of an example given by Woodward (2003, 94-95).
[6] As an anonymous reviewer pointed out, Woodward's views on the role and nature of interventionism have considerably evolved over the years. In earlier writings, Woodward presented interventionism as at least partly a conceptual or semantic project (e.g., Woodward 2003, 38), but recently he has characterized it explicitly as a methodological or functional account that should be judged by its usefulness (Woodward 2014, 2015). I follow here this more recent understanding of the theory.

It seems to capture well the role of causal thinking in fields such as biology, economy, or psychology. Thus, interventionism constitutes an independently plausible and fruitful approach to causation, and even for those skeptical of interventionism, it should be interesting to find out what its implications are for this issue.

Recently many authors have argued that interventionism allows for genuine causal roles for psychological states (e.g., Eronen 2012; Menzies 2008; Raatikainen 2010; Shapiro 2010, 2012; Shapiro and Sober 2007; Weslake forthcoming; Woodward 2008a, 2015a). The idea is that there are possible interventions on psychological states that result in changes in behavior when all other relevant variables are held fixed to their values, and thus psychological states are causes of behavior. Moreover, these causes are not excluded by any possible neural causes of behavior, but can peacefully coexist with them: The fact that there are both psychological and neural causes of behavior merely means that there are counterfactual patterns of dependency both between psychological states and behavior and neural states and behavior.

These arguments have been extensively discussed elsewhere; here I will focus on how this reasoning can be extended to Dennettian intentional states. In fact, the framework of the intentional strategy is exceptionally well-suited for an interventionist treatment. To see why, let us first return to the example in section 2. When Jim utters "I am going to get a cup of coffee now", Jane can fairly reliably predict that he will go to the kitchen, based on attributing to him the belief that there is coffee in the kitchen and the desire for coffee. However, the intentional strategy is not just a brute predictive tool. The predictions work because behavior depends on beliefs and desires in the sense that variation in beliefs and desires is reliably associated with variation in behavior. In the case of Jim, the belief that there is coffee in the kitchen combined with a desire for coffee is associated with the behavior of going to the kitchen. Thus, there are real patterns in human behavior, and these patterns form the basis for predicting and explaining human behavior (Dennett 1987, 27; 1991; see also Richard 1994). These patterns are in general stable and objective enough to support generalizations and predictions (Dennett 1987, 25). Moreover, if they allow for voluminously successful prediction, they must also tell what would happen under different conditions that were not actually realized. In order to be a good predictor, Jane must not only be able to predict that Jim goes to the kitchen if he believes that there is coffee there, but also that if Jim believes that the coffee is finished, then he probably goes straight to the cafeteria instead. Thus, the patterns underlying human behavior support counterfactuals and can be relied upon when seeking answers to what Woodward (2003) calls "what-if-things-had-been-different" questions. In this light, it seems very plausible that intervening on Jim's belief concerning coffee in the kitchen would be one way of changing Jim's behavior.

Moreover, scientific practice suggests that researchers often take something like the intentional stance on human subjects and then successfully isolate intentional states as interventionist causes for behavior. Let us take misinformation experiments as an example (Loftus et al. 1978, Loftus 2005). In these experiments, subjects are first shown an event, and then after a delay, are given partly misleading information about the event, which typically affects the memory of the event. For example, the subjects are shown a video of a pickpocket with a black hat stealing a wallet. After a delay, they receive information that the pickpocket had a red hat. When asked about the event, they now report that the pickpocket had a red hat. Misinformation experiments of this kind clearly involve interventions that target specific beliefs of the subjects (i.e., their beliefs about the event), and changes in those beliefs result in changes in the (verbal) behavior of the subjects. If the randomization in the trial is done correctly, the test and control groups are large enough, and other standard protocols of experimental design are followed, then possible confounding factors can be controlled for, and it is likely that the experimental effect really is due to the intervention, i.e., the misinformation given.

Thus, it seems that the interventionist treatment of mental causation can be easily extended to states ascribed from the intentional stance, and that this is also supported by scientific practice. I will discuss objections and remaining problems in section 5. At this point, it is important to emphasize that the position I am defending in this paper does not require that the success of the intentional strategy somehow *implies* interventionist mental causation. The main problem for the intentional stance as a philosophical position has been that it is not clear how Dennettian intentional states even *could* be causes. In other words, the position seems to be conceptually inconsistent, as it is supposed to amount to a form of intentional state realism, but does not seem to leave any genuine causal role for intentional states. What I aim to show is that there is an interventionism-based version of the account where intentional states *can* be genuine causes. However, is interventionist mental causation at all compatible with the main tenets of the intentional stance account? I will now turn to answering this question.

4. The Reality of Intentional States

Above I have pointed out that interventionism provides a way of making sense of how intentional states ascribed from the intentional stance can be real causes. However, this may seem to undermine the most appealing feature of the intentional strategy, namely that it makes beliefs and desires objective and real phenomena while avoiding strong realism. Taking intentional states to be real causes may seem to imply that they are real in a rather strong sense.

However, it is important to understand that the kind of realism we want to avoid here is what Dennett (1991) calls "industrial strength realism" and attributes to Fodor, and that is arguably common among contemporary neuroscientists. On this view, beliefs and desires are states of the brain whose presence or absence we could in principle objectively verify by looking into the brain in the right way. For example, to hold the belief that p is to have a token with the content p in the "belief box" (or something similar), and if we could look into the brain at the right level, we could check whether someone has that belief or not.

Although treating intentional states as interventionist causes does imply that they are real, it does not lead to this kind of realism. Interventionism only requires that there is a well-defined and coherent conception of how the putative causes can be changed or manipulated, and does not put any further constraints on the ontological nature of causes (Woodward 2003, 111-114). Examples of interventionist causation include changes in the editorial policy of a newspaper causing changes in voting behavior, a freeze in Florida causing a rise in the price of oranges, differences in hospitalization regimes having a causal effect on patient recovery, and so on (Woodward 2003; 2015). It is clear that interventionism does not imply that intentional states are objectively verifiable states or structures (such as computational or neurobiological states) in the brain, or something analogous to tokens in a belief box. The fact that intentional states are causes of behavior only implies that it is possible to intervene on them in the right way in order to change behavior (see section 3).

Thus, adding interventionism into the picture does not collapse the intentional stance into the kind of strong realism that takes intentional states to be objectively verifiable things in the brain. Nevertheless, some proponents of a Dennettian approach may still find the view defended here too strong. One way of reading Dennett is that he denies that beliefs are any kinds of things at all: Only the (behavioral) patterns really exist, while beliefs are merely something that we ascribe to agents to predict their behavior (cf. Dennett 1987, 37-42). This clearly conflicts with the view that beliefs are real interventionist causes, as in this picture only patterns are real, not the beliefs themselves. The problem with this kind of eliminative pattern-realism, however, is that if only the behavioral patterns really exist, then it seems impossible to account for the causal role of beliefs and other intentional states. As it is a fact that things that do not exist cannot be causes of anything, this position leads again to the kind of instrumentalism that most readers of Dennett have found unacceptable. The appeal of the interventionist approach defended here is that it grants beliefs a causal role in a way that is compatible with the central Dennettian thesis that they are not objectively verifiable things in the brain.

Thus, the realism that the interventionist approach implies is not too strong – it is arguably the weakest kind of realism that you can have and still hold on to intentional state causation. However, more realistically oriented philosophers of mind may wonder whether this realism is in fact *too weak*. If intentional states are only discernible from the intentional stance (see section 2), and interventionism is a very metaphysically minimalistic theory of causation, in what sense are intentional states supposed to be real?

First of all, although interventionism may appear to be an entirely context-relative and ontologically non-committal account of causation, this is not the case. If X is a cause of Y in some representation of system S, then X is a cause of Y in all representations of system S where X and Y appear. This follows from the definition of an intervention, which is entirely general and not relativized to a variable set (see Woodward, 2008b for more). Furthermore, causal relationships in interventionism are based on objective relationships of counterfactual dependency that are mind- and interest-independent. Once the variables are selected, these patterns of counterfactual dependency make causal claims true or false in an entirely objective way (Woodward 2003, 118-122; see next section for complications related to supervenience). For these reasons, interventionism can be seen as a realist account of causation, and interventionist causes can be seen as real causes (ibid.).

Secondly, there are many phenomena in the special sciences that are analogous to Dennettian intentional states in the following sense: Due to the immense complexity of their physical basis, they can only be discerned from a certain perspective (or scale), but we nevertheless have very good reasons to consider them to be real. A good example is climate teleconnections, discussed by Glymour (2007).[7] These teleconnections are stable relationships between aggregate temperature measurements of different regions. For example, the Northern Atlantic Oscillation (NAO) is characterized by fluctuations between the Icelandic low atmospheric pressure center and the Azoric high atmospheric pressure center (Hurrell 1995; Hurrell et al., 2003). Teleconnections such as NAO are considered to be real and objective phenomena by scientists (ibid.), and as Glymour (2007, 340-342) points out, they can be seen as causally relevant factors in explaining variations in climate. However, although they have a physical basis in the physical features of the earth and the atmosphere, and ultimately in the movements of particles, their relationship to that basis is intractably complex. More or less similar examples can be found in neuroscience (e.g., event-related potentials, attractors in dynamic models) and in other fields of science, such as economics (e.g., monetary transactions) or physics (e.g., various cases of universality, see Batterman 2002).

---

[7] Dennett himself compares beliefs to centers of gravity, but in my view this is example is not very helpful, as the analogy to intentional states is rather weak: Dennettian intentional states are not supposed to be related to the brain or other physical stuff in any straightforward way, while the physical basis of centers of gravity is relatively obvious and unproblematic.

Why should we consider phenomena like these to be real? There are many answers to this in the literature (e.g, Batterman 2002; Ladyman and Ross 2007; Ross 2000; Ross and Spurrett 2004; Viger 2000), but what I consider to be the strongest argument is the following: We are justified in believing that these phenomena are real insofar as they are *robust*, i.e., insofar as they can be measured, modeled, detected, produced, or the like, in a variety of independent ways (Eronen 2015; Wimsatt 1981, 2007). The basic idea is that if there are several such independent ways, then the likelihood that they all happen to turn out to be mistaken is extremely low, and we have very good reasons to think that the phenomenon is real (see Eronen 2015 for a detailed discussion of this argument). It is clear that NAO is robust in this sense: It can be detected from a broad range of independent temperature measures and appears in various independent models of the climate (Hurrell 1995; Hurrell et al., 2003). Similarly, we can be justified in believing that intentional states are real insofar as they are robust.

This may seem to be in conflict with Dennett's claim that the only way to ascertain whether someone has beliefs or desires is to take the intentional stance (see Section 2). However, what this means is that intentional states and the patterns underlying them are only discernible at a higher behavioral scale, and not, e.g., from a microphysical or neurobiological perspective (Dennett 1991, 1987, 28). These higher-scale phenomena and patterns are still discernible and verifiable by independent methods and observers, and therefore can be robust. If Dennett (1987, 2009) is right, intentional states play an important role in models and explanations in many different branches of sciences, and are even potentially discernible by Martians (if they focus on the right scale), and are thus very robust.

One might object that even though intentional states and other similar phenomena may be real in the weak sense of being robust higher-scale phenomena, they are not real in a metaphysically *deep* sense, e.g., in the sense of having distinct causal powers (Kim 2005) or being composed of microphysical particles and governed by physical regularities (Pettit 1993). However, if these metaphysically deep criteria for what is real have the implication that a broad range of things that are studied in science and treated as real by scientists turn out to be unreal, this is a good reason to doubt the validity of such criteria. Furthermore, even if we accept the importance of such criteria for metaphysical debates, it can plausibly be argued that they are irrelevant here. The issue is not the final ontological make-up of the world or the general relationship between special science things and fundamental physical things, but whether we are justified in considering Dennettian intentional states to be real in the sense that other special science phenomena are real. Above I have argued that insofar as they are robust, they are analogous to phenomena such as climate teleconnections or monetary transactions, and therefore should also be considered to be real in the same way.

The considerations of this section and section 3, when taken together, result in a novel ontological picture. If we can voluminously predict the behavior of a system by taking the intentional stance, then this is strong evidence that the system has intentional states. Furthermore, as I have argued above, if those intentional states and the patterns underlying them are robust, then we are justified in believing that they are real, even if they are not (identical to) states of the brain. The most state-of-the-art version of interventionism can then be applied to show how such intentional states can have a genuine role in causing behavior, as I will explain in detail in the next section.

In this picture, we do not need to show that intentional states are in some specific sense reducible to physical states in order to conclude that they are real causes (contra Kim 2005), and we do not need to show that they are in some specific sense physically realized in order to conclude that they are real and ontologically acceptable. Instead, we are justified in considering intentional states to be real causes if they are (1) voluminously predictive; (2) robust; and (3) satisfy the conditions of interventionism.

It is of course far from clear whether intentional states actually are like this: How successful is intentional stance prediction in the end? How robust are intentional states and the higher-scale patterns underlying them? Can't we still find something like intentional states or their correlates in the brain? Perhaps some other conceptual framework will turn out to be more robust and predictive for human behavior? However, these are empirical questions that need to be empirically settled. What I have argued above is that it is a distinct and conceptually consistent possibility that intentional states are of this nature. I will now turn to remaining problems with the interventionist treatment of intentional states.

5. Interventionist Exclusion Worries[8]

Perhaps the most serious obstacle to treating intentional states as interventionist causes is the so-called interventionist causal exclusion problem, which arises from the following exclusion argument (Baumgartner 2009b).[9] It is widely accepted that psychological states supervene on neural (or physical) states, which implies that changes in psychological states are always accompanied by

---

[8] I am grateful to an anonymous reviewer whose insightful comments provided the impetus to write this section.
[9] Recently Gebharter (forthcoming) has presented a sophisticated argument in the same vein, purporting to show that the causal exclusion argument works in the framework of causal Bayes nets. The response in this section can be adapted to that framework as well; this will be addressed in the detail it deserves in future publications.

changes in some neural (or physical) states.[10] One explanation for such supervenience is that psychological states are simply identical to neural (or physical) states. However, the picture defended above requires that intentional states *non-reductively* supervene on neural (or physical) states. If this is the case, how can we intervene on a psychological variable while holding all other variables (that are not on the path from the putative cause to the effect) fixed? Supervenience implies that it is impossible to hold all neural (or physical) variables fixed, but holding all other variables fixed is a necessary condition for interventionist causation, so this seems to lead to the conclusion that interventionist psychological causation is impossible (see Baumgartner (2009b) for more details).

A straightforward solution to this problem is to drop the requirement of holding *all* other variables fixed, and to allow for changes in variables that are correlated with the variables of interest due to supervenience, definition, composition, or some other non-causal relationship (Eronen & Brooks 2014; Weslake, forthcoming; Woodward 2015a; see also Campbell 2007). To allow for this, the definitions of causation and intervention have to be slightly modified, roughly as follows (Baumgartner 2013; Woodward 2015a): X is a cause of Y (in variable set **V**) iff the value (or the probability distribution) of Y changes under some intervention on X, when all other variables in **V** (that are not on the path from X to Y) are held fixed, *except for those variables that are related to X or Y (or the variables on the path form X to Y) by supervenience (or definition, or composition, etc.).* An intervention on X with respect to Y has to cause the change in X; the change in X has to be entirely due to the intervention and not any other factors; the intervention should not change Y directly; and it should be uncorrelated with any causes of Y that are not on the path from X to Y, *except for those causes of Y that are related to X by supervenience (or definition, or composition, etc.).*

This may seem ad hoc at first, but non-causal relationships (supervenience, composition, conceptual relationships, etc.) are also ubiquitous in sciences such as neuroscience, biology, or chemistry, and we need a way of dealing with such relationships, without eliminating or reducing all higher-level causes. In practice, scientists do not require holding the supervenience base fixed when assessing the causal role of the supervenient variable (Woodward 2015a). Otherwise one would have to conclude that monetary transactions can have no causal relevance unless they are identical to some lower-level variable(s), the editorial policy of a newspaper can have no causal relevance unless it is identical to some lower-level variable(s), and climate phenomena such as NAOs have no causal relevance unless they are identical to some lower-level variable(s). In short, all special sciences variables would

---

[10] It is not clear whether Dennett thinks that intentional states supervene on neural states; one interpretation is that he thinks that they supervene on behavior instead, and the issue of supervenience and the original intentional stance is extremely convoluted (see McLaughlin and O'Leary 1994 for more). Following the broad consensus in philosophy of mind, I assume here that some kind of mental-to-physical supervenience holds for intentional states.

have to be either identical to some lower-level variable(s), or else epiphenomenal. This would run counter to scientific practice, and undermine the whole rationale of interventionism, so the revision where the supervenience base is allowed to vary is in fact well-motivated independently of the issue of psychological causation.

However, the revised version of interventionism faces a further problem: It seems to make mental causation a matter of stipulation, as epiphenomenal causal structures become impossible, or at least empirically indistinguishable from corresponding non-epiphenomenal structures (Baumgartner 2013). To see this, let us consider a situation where M non-reductively supervenes on P, and P is a cause of B. Due to supervenience, it seems that it will always be possible to intervene on M with respect to B: One can intervene on M so that P changes and causes a change in B, and in the revised version, this kind of an intervention counts as an intervention on M with respect to B (see Baumgartner 2013 and Eronen & Brooks 2014 for more detailed explanations). Thus, it seems that M will *always* be the cause of B in structures like these, or more generally, a mental state will always be a cause of the effects of its supervenience base (Baumgartner 2013). Moreover, the difference-making relations and correlations in the structure where M *is* a cause of B will be exactly the same as those of the same structure where M is *not* a cause of B. In other words, the epiphenomenal causal structure is empirically indistinguishable from the non-epiphenomenal one. This in turn suggests that the revised version of interventionism makes the existence of mental causation an entirely non-empirical issue that is solved by simply stipulating that there are no epiphenomenal causal structures.

In order to respond to this, we need to analyze the setting in more detail. Note that in the standard representation of mental causation that is also the basis of Baumgartner's argument there is only one variable P in the supervenience base of M. In reality, the supervenience base can also include many different variables. Moreover, as multiple realizability is one of the core assumptions of non-reductive physicalism, we should allow for the possibility that M can be realized by several different sets of variables, or by several different combinations of values of the variables in the supervenience base. This is certainly the case with the kinds of intentional states discussed in this article, as one of the core assumption of the intentional stance is that the physical basis of intentional states is immensely complex.

One possible way of representing this kind of multiple realizability is as follows: M (the intentional state) supervenes on a large set of physical variables P_i. Certain combinations of values of those variables correspond to one value of M (say, M=1), and certain other combinations of values correspond to another value of M (say, M=0). M is a putative cause of a behavioral effect B.

In this framework, the implications of revised interventionism are far more complex and nuanced than in the simple standard diagram. If we intervene on M with respect to B, at least one of the physical variables, say P_x, must change in value, but which P_i variable this is can differ from one occasion or context to the next. It is also possible that P_x can only change when other P_i variables are set to certain values, or that there are other background conditions (e.g., values of P_i) where changes in P_x are not associated with changes in B. Moreover, there will be ways of intervening on M with respect to B so that P_x does *not* change (but rather some other P_i changes). Thus, even though P_x will be a cause of B, this causal relationship can be very shallow and uninteresting, and will certainly be different from the more stable and general causal relationship between M and B.[11]

For similar reasons, in a complex structure like this, it is possible that, contra Baumgartner, M is not a cause of all the effects of its supervenience base. For example, it could be that one can intervene on P_x with respect to some effect B *only* in background conditions C (e.g., with other P_i variables set to certain specific values), and that in those background conditions, changing M is not a possible way of changing P_x. In this setting, P_x is a cause of B, but M is not. For the same reason, it is also possible that M supervenes on physical variables, but is nevertheless epiphenomenal with respect to physical behavior: Changes in M would then result in changes in the supervenience base, but somehow never in the right combinations of values that would result in a P_i causing B.[12]

More importantly, in this picture intentional state causation is not just a matter of non-empirical stipulation. In the original diagram where M supervenes on only one P variable, it may seem arbitrary to treat M as causally relevant over and above P, and the status of M as a distinct variable or property can be questioned. However, in the picture where M is multiple realized by many combinations of variables, the relationship between M and B seems to capture important additional causal information that can also be exploited for purposes of manipulations and control. The relationships between the P_i variables and B can be unstable and context-dependent, whereas the relationship between M and B is more stable and general. Moreover, whether we can find higher-scale variables like M that efficiently capture lower level variation is an empirical question: it is also possible that in some systems causal relationship among physical variables do not give rise to any interesting higher-scale patterns.

It is true that once we have figured out the full causal structure and the exact relationships between higher- and lower-level variables there is no further empirical test that could confirm or disconfirm

---

[11] Campbell (2010) calls variables such as M "control variables", and Woodward (2008a) refers to structures of this kind as involving "realization-independent dependency relations".
[12] It should also be noted that Baumgartner's argument only applies to epiphenomenalism due to supervenience: Causal structures where the epiphenomenal variable is simply the end point of a causal chain are still perfectly possible.

mental causation, and it follows then from interventionism that certain types structures exhibit higher-level or downward causation. However, this does not make the issue as a whole non-empirical. There remains a very substantial empirical part in determining whether there is mental causation, namely figuring out the causal structure in the mind-brain and the relationships between variables at different levels. The interventionist approach to higher-level causation does not collapse into a trivial or non-empirical solution.

6. Concluding remarks

In this paper, I have shown that the intentional stance, when combined with interventionism, is a plausible and consistent position on the nature of intentional states and their relationship to the brain. However, as pointed out in the introduction, the position does not incorporate *all* of the aspects of Dennett's (1987) intentional stance theory, such as its behavioristic or Rylean aspirations. To distinguish the two, the present account could be called the *interventionist stance*.

Connecting intentional systems theory with interventionism opens up new avenues of research. The fact that there is an interventionist causal relationship between two variables is only minimally informative and the starting point of inquiry. Further questions that can be asked include (among many others; see Woodward 2010): How stable is that relationship? Stability here refers to the degree to which the relationship continues to hold in various circumstances. How specific is the relationship? Specificity refers to the degree to which the relationship exhausts the causal relationships of the relata, i.e., if the relationship is maximally specific, then X is the only cause of Y and X has no other effects than Y. In this way, bringing interventionism into the picture allows us to not only conclude that intentional states can be causes, but also to analyze the causal relationships in more detail. For example, it is quite plausible that if intentional states are like described in this paper, then then relationships between intentional states and behavior are far more stable than any relationships between the physical variables in the supervenience base and behavior. Regarding specificity, due to the supervenience argument intentional states cannot be maximally specific, but they are nevertheless likely to be far more specific causes of behavior than the physical causes in the supervenience base.

One possible worry that remains is that this interventionist stance may lead to the conclusion that systems such as thermostats, neurons and chess-playing computers have beliefs that are real causes, insofar as we can voluminously predict their behavior from the intentional stance. This is something

Dennett (2009, 87-88) would probably accept, but most other philosophers would find highly implausible. However, in interventionism there has to be some reasonably clear sense of what an intervention on a variable would involve, and how it would change the value of the variable (although it does not need to be possible in practice) – for example, there is no clear sense in which we could intervene to change the species of an organism (Woodward 2003, 111-114). In a similar way, it can be argued that we have no clear conception of how we could intervene on the "beliefs" of a neuron or a thermostat. Chess-playing computers are more difficult cases. One possible response is to bite the bullet and to argue that if we can indeed voluminously predict the behavior of such computers from the intentional stance, and successfully manipulate their behavior from this stance, then there is a sense in which such computers have beliefs that are real causes. Another option that I prefer is to appeal to robustness: It is unlikely that the beliefs in chess-playing computers are detectable in many independent ways, so those beliefs will at best be robust to a low degree only.

Recently Ladyman, Ross and Collier[13] have put forward a spirited defense of the reality and objectivity of higher-scale patterns that also supports the position defended here (Ladyman and Ross 2007, ch. 4). Building on Dennett's idea of real patterns, Ladyman, Ross and Collier present an elaborate definition of real patterns based on the concepts of information, encoding and projectibility. They also argue for the "scale-relativity of ontology", the idea being that claims about what is real are relative to the scale at which things are measured. In their picture, higher-level phenomena such as intentional states can be seen as real patterns that are discernible only at certain scales and irreducible to patterns at lower scales. However, in contrast to what I have defended in this paper, Ladyman and Ross (2007, ch. 5) take causation to be just a heuristic for tracking real patterns, and do not subscribe to a difference-making or interventionist account of causation. They argue that causation should be seen as folk notion that does not have a fundamental role in scientific explanations or theories (ibid., see also Ross and Spurrett 2004). This is problematic, and can be seen as a shortcoming of the account, as causal explanation and the search for causes does seem to be centrally important for most special sciences, and special science causes are widely regarded as real causes.

Although I do not want to defend the extreme form of a real-pattern ontology of Ladyman, Ross and Collier, it is worth considering how we could integrate interventionist mental causation (or interventionism in general) into this framework as well. This could be done roughly as follows. Taking intentional states as an example, there are two types of patterns at work here: Firstly, there are the intentional states themselves, which in the patterns-ontology of Ladyman, Ross and Collier can be

---

[13] John Collier is the third author of the chapter that I discuss here (Ladyman and Ross, 2007, ch. 4), but is not listed as a main author of the book.

interpreted as real patterns. Secondly, there are patterns of counterfactual dependency between the intentional states and behavior (or other intentional states). These latter ones are the kinds of patterns that underlie interventionist causal relationships (see above). Taken together, these two patterns result in interventionist intentional state causation. Both kinds of patterns are real and only discernible at the right scale. In this way, we could incorporate special science causation into the real-patterns ontology, without making causation just a heuristic tool or a folk notion. This would arguably make the real-patterns approach in the vein of Ladyman, Ross and Collier more plausible and attractive, and illustrates how one could embed interventionist intentional causation even into a purely pattern-based ontological framework.

In conclusion, the account given in this paper vindicates the core features of the intentional strategy, and also connects it to various other topics in contemporary philosophy of science, opening the door to further inquiries and possibilities. Far from being outdated and peripheral, the intentional stance in this form is a viable theory of the nature of intentional states and their relationship to the brain.

Acknowledgments

References

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

Baumgartner, M. (2009a). Interdefining Causation and Intervention. *Dialectica* 63, 175-194

Baumgartner, M. (2009b). Interventionist Causal Exclusion and Non-reductive Physicalism. *International Studies in the Philosophy of Science*, 23, 161-178.

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica*, 67, 1-27.

Campbell, J. (2007). An Interventionist Approach to Causation in Psychology. In A. Gopnik and L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy and Computation* (pp. 58-66). Oxford: Oxford University Press.

Campbell, J. (2010). Control Variables and Mental Causation. *Proceedings of the Aristotelian Society*, 110, 15-30.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60, 685-700.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68, 87-106.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, 88, 27-51.

Dennett, D. C. (1996). *Kinds of minds. Towards an understanding of consciousness.* New York: Basic Books.

Dennett, D. C. (2000). With a little help from my friends. In D. Ross, A. Brook & D. Thompson (Eds.) *Dennett's philosophy. A comprehensive assessment* (pp. 327-388). Cambridge, MA: MIT Press.

Dennett, D. C. (2009). Intentional systems theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.) *The Oxford Handbook of Philosophy of Mind* (pp. 339-350). Oxford: Oxford University Press.

Eronen, M. I. (2012). Pluralistic Physicalism and the Causal Exclusion Argument. *European Journal for Philosophy of Science* 2(2), 219-232.

Eronen, M. I. (2015). Robustness and Reality. *Synthese* 192(12), 3961-3977.

Eronen, M. I. & Brooks, D. (2014). Interventionism and Supervenience: A New Problem and Provisional Solution. *International Studies in the Philosophy of Science* 28(2), 185-202.

Gallagher, H. L., Jack, A. I., Roepstorff, A. and Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, 16, 814-821.

Gebharter, A. (forthcoming). Causal exclusion and causal Bayes nets. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12247

Gergely, G., Nádasdy, Z., Csibra, G. and Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165-193.

Glymour, C. (2007). When is a brain like the planet? *Philosophy of Science*, 74, 330-347.

Griffin, R. and Baron-Cohen, S. (2002). The intentional stance: developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.) *Daniel Dennett* (pp. 83-115). Cambridge: Cambridge University Press.

Horgan, T. and Woodward, J. (1985). Folk psychology is here to stay. *The Philosophical Review*, 94, 197-226.

Hurrel, J. W. (1995). Decadal trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, 269, 676-679.

Hurrell, J.W., Kushnir, Y., Ottersen, G. and Visbeck, M. (2003). An overview of the North Atlantic Oscillation. In J. W. Hurrell, Y. Kushnir, G. Ottersen, & M. Visbeck (Eds.) *The North Atlantic Oscillation. Climatic significance and environmental impact. Geophysical Monograph 134* (pp. 1-35). Washington, DC: American Geophysical Union.

Jackson, F. and Pettit, P. (1990). Program explanation: a general perspective. *Analysis*, 50, 107-117.

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

Ladyman, J., and Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalised*. Oxford: Oxford University Press.

Lewis, D. (1973). Causation. *The Journal of Philosophy* 70, 556–67.

Loftus, E. F., Miller, D. G., and Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of experimental psychology: Human learning and memory*, 4, 19-31.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12, 361-366.

McCulloch, G. (1990). Dennett's little grains of salt. *The Philosophical Quarterly*, 40, 1–12.

McLaughlin, B. P. and O'Leary-Hawthorne, J. (1994). Dennett's logical behaviorism. *Philosophical Topics*, 22, 189-258.

Menzies, P. (2008). The exclusion problem, the determination relation, and contrastive causation. In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pettit, P. (1993). A definition of physicalism. *Analysis* 53, 213-223.

Pöyhönen, S. (2014). Intentional concepts in cognitive neuroscience. *Philosophical Explorations*, 17, 93-109.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis*, 73, 349-363.

Rey, G. (1994). Dennett's unrealistic psychology. *Philosophical Topics*, 22, 259-289.

Richard, M. (1994). What isn't a belief? *Philosophical Topics*, 22, 291-318.

Ross, D. (2000). Rainforest realism: A Dennettian theory of existence. In D. Ross, A. Brook & D. Thompson (Eds.) *Dennett's philosophy. A comprehensive assessment* (pp. 147-168). Cambridge, MA: MIT Press.

Ross, D. and Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences* 27, 603-647.

Shapiro, L. (2010). Lessons from causal exclusion. *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, L. (2012). Mental manipulations and the problem of causal exclusion. *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, L. A. and Sober, E. (2007). Epiphenomenalism: the dos and the don'ts. In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Slors, M. V. P. (2007). Intentional systems theory, mental causation and empathic resonance. *Erkenntnis*, 67, 321-336.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction, and search*. New York: Springer.

Viger, C. (2000). Where do Dennett's stances stand? Explaining our kind of mind. In D. Ross, A. Brook & D. Thompson (Eds.) *Dennett's philosophy. A comprehensive assessment* (pp. 131-145). Cambridge, MA: MIT Press.

Wimsatt, W. C. (2007) *Re-engineering philosophy for limited beings. Piecewise approximations to reality*. Cambridge, MA: Harvard University.

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

Woodward, J. (2008a). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation*. Oxford: Oxford University Press: 218-262

Woodward, J. (2008b). Response to Strevens. *Philosophy and Phenomenological Research*, 77, 193–212.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy* 25, 287-318.

Woodward, J. (2015a). Interventionism and causal exclusion. *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, J. (2015b). Methodology, ontology, and interventionism. *Synthese* 192, 3577-3599.

Zawidzki, T. W. (2012). Unlikely allies: embodied social cognition and the intentional stance. *Phenomenology and the Cognitive Sciences*, 11, 487-506.