Investigating Causal Effects of Mental Events in Cognitive Neuroscience

Mikkel C. Vinding, Ph.D.

NatMEG, Department of Clinical Neuroscience
Karolinska Institutet
Nobels väg 9, D3
171 77 Stockholm
Sweden

Email: mikkel.vinding@ki.se

Abstract

Mental causation is a predominantly theoretical topic rather than a topic studied in the laboratory. The purpose of this paper is to outline a general approach for studying mental causation by empirical means for philosophers and scientists interested in the topic. The aim is to outline how we can infer mental causation by empirical methods given an unknown solution to the mind-body problem. The approach is based on the principles of causal inference to find causal relations among observed variables used in all branches of science. With these principles, it is possible to estimate the causal effects of mental events: Make an experimental manipulation on a mental event, control confounding variables, and estimate causal effects on the outcome. The caveat is that we cannot separate the causal effects of a mental event from the physical base of the mental event, independent of whether we assume mental events can be reduced to their physical base. A challenge to estimating causal effects of mental events is that measured physical variables, such as electrophysiological potentials from the brain, might reflect processes that are part of "higher-order" phenomena, such as mental events. This means that controlling "downwards" for confounding variables is challenging. It is, however, not impossible. It also means that inferring non-mental causes of action cannot be done by measuring only physical variables alone. Keeping the mind-body problem in mind when designing experiments, it is possible to infer mental causation.

Keywords: Mental Causation, Causal Inference, Cognitive Neuroscience, Mental Events, Mind-Body Problem

1 Introduction

Do mental events cause physical action, and if so, how? This is the central question in the topic of mental causation. One ought to think that investigating mental causation is a goal of experimental psychology—but experiments directly addressing mental causation are surprisingly sparse. Even though a massive advancement in neuroscience methods has given experimental psychology and neuroscience tools to study the biological basis of mental events, there has been little attempt at a framework for empirical investigations of mental causation. Mental causation is predominantly a theoretical topic with close to no contribution from experimental science.

The central problem in addressing mental causation is the mind-body problem. The mind-body problem, in a nutshell, is that we do not know how mental events relates to physical states. Without knowing how they are related, we do not know how they interact. The mind-body problem makes it difficult to define the pre-empirical foundation for experimental inference about mental causation. We assume the brain is the foundation of the mind, but since we do not know how they are related, it is unclear what assumptions we have to make to include mental events when inferring causal relations. The mind-body problem appears to be an obstacle to describing the foundation for investigating mental causation by empirical means.

The purpose of this paper is to sketch a foundation for an experimental approach that scientists and empirical oriented philosophers can use to study mental causation. How do we go from the analytical approach to understanding mental causation to instead gain an understanding via experimental inference? Furthermore, how do we distinguish between the type of questions we must deal with by analytical reasoning and those questions we can answer by empirical means?

Given the unknown relationship between mental events and the physical world—the mind-body problem—I will explore to what extent we can make meaningful inference about mental causation. As we must be aware of what we can (and cannot) infer from experimental studies on mental causation, I will make the necessary assumptions explicit and acknowledge limitations in the experimental designs. Finally, I will answer the following questions: what type of questions about mental causation can we answer through experimental procedures? And what possible caveats must we avoid to draw the right conclusions from experimental studies?

In the following sections, I discuss how we deal with mental and physical phenomena as experimental variables and outline how we can use them to make inference about causal relations, how the variables relate to analytical problems, and how this relation sets the foundation for an empirical approach to studying mental causation.

2 Mental variables and physical variables

The analytical approach to studying mental causation focus on how any ideal mental event *M* and any ideal physical event *P* can (or cannot) interact. In contrast, empirical science deals with observed data to infer relations amongst the events the variables represent. When taking an empirical approach, our first assumption is that mental and physical events are real and that we can measure and/or manipulate them. To measure and manipulate mental event, they need to be operationalized as experimental variables.

The first hurdle is that the term *mental event* has different meanings in different discussions. It can refer to specific mental content occurring within a limited time window or refer to general states, e.g., a transient intention to move one's arm versus a general state of wakefulness or being in a coma (Hohwy, 2009; Laureys, 2005). There is no clear-cut definition of mental events; partly due to the uncertainty in defining the nature of mentality, to begin with. The content of mental events can refer to phenomenological properties or cognitive properties (Block, 2005; Cohen and Dennett, 2011). The precise definition of mental events is not of importance: as long as we can accept that mental events exist—either as phenomenological states or cognitive processes—then we can operationalize these as *mental variables* in experimental settings. Thus, mental events can be defined by their phenomenological content, or they can be defined from a cognitive perspective without referring to phenomenology.

If we are strict, we could argue that because the phenomenological content of mental events is available only to the subject, we can never measure them. We can for example never know if subjects have inverted qualia, or if they are philosophical zombies (Chalmers, 1997). There is no practical solution to this problem, but this does not exclude mental variables from being meaningful in experimental contexts (Overgaard et al., 2008; Seth et al., 2005; Tononi and Koch, 2015).

Subjective reports or behavioral responses might not be "direct" access to phenomenological content, but they serve as indirect indications that mental events are occurring. We see indirect variables in all branches of science: when measuring distortion of light as an indication of cosmic bodies due to gravitational bends, and in neuroimaging where the blood-oxygen flow in areas of the brain is a proxy of neuronal activity (Logothetis et al., 2001). We accept these indirect measurements because the link between the indirect measure (light distortion; oxygen-blood flow) and the object (cosmic body; neural activity) is based on assumptions that we agree upon (strong gravitation bends light; active neurons are associated with higher blood flow).

To obtain a mental variable that indicates that a subject is experiencing a particular mental event is sufficient for it to be an experimental variable. We can measure mental events, by obtaining introspective reports about the mental events or other indications that subjects are experiencing certain mental events (Overgaard et al., 2008). What we obtain by these methods are *mental variables.* For mental variables to be valid, they have to be consistent as any other variable. Mental variables should exclusively capture the event they are intended to measure while exhaustively capturing any occurrence of the mental event (Jensen et al., 2017; Reingold and Merikle, 1988).

We are not required to know the ontological reality of the mental events that the mental variables measure, i.e., we do not need to impose a predefined solution to the mind-body problem to use mental variables in cognitive neuroscience. If we can trust our methods of obtaining mental variables, then we have indications that the mental events are occurring, and we are justified in using these as variables in experimental studies.

Physical events seem more intuitive than mental events, but, upon further inquiry, it is not straightforward what constitutes a physical event (Crane and Mellor, 1990; Melnyk, 1997; Smart, 1978). Intuitively, we can easily characterize different events, such as a neuron firing, cerebral blood-flow, or the force of an accelerating mass as physical events. Although only the latter is the described in the scientific language of physics, we consider all previous examples as physical events. That we consider the above as physical phenomena are because we not have any reason to assume their existence is dependent on anything that violates the language of physics (Stoljar, 2001). In this sense, cognitive and neural processes—from single neurons to whole-brain network communications—are all

physical phenomena. Whatever we measure in cognitive neuroscience that relates back to the biophysical properties of the brain or body is in this context a *physical variable*.

The challenge of using an empirical approach to studying mental causation is how to combine mental variables and physical variables in experiments to infer their causal relation. Another factor that adds to this challenge is that there are likely as many definitions of *causation* as there are definitions of *mental events*. Despite disagreements, there is one prevailing view of how to make causal inference in science. In the next section, I give a brief overview of the general principles of causal inference and then return to how we can use these principles to make inference about mental causation.

3 The principles of causal inference

The causal effect of X is the difference X would make on the outcome Y, based on the counterfactual conditions of X being present or not. Say we want to know whether X (a drug) cause Y (a fever reduction). Depending on the intervention the variable X takes the values X=1 (taking the medication) or X=0 (not taking the medication). Y is the measured outcome of the counterfactual conditions, as $Y(X=1)$ and $Y(X=0)$. If X is causing Y, then Y will follow X=1 but not X=0. Thus, the *causal effect* of X on Y is the measured difference between $Y(X=1)$ and $Y(X=0)$. The causal effect is a numerical quantity that indicates the difference between the counterfactual conditions of X and not X (Rubin, 1974; Woodward, 2005).

In reality, both conditions cannot occur: a subject cannot take the pill (X=1) and at the same time not take the pill (X=0). Only one of the counterfactual conditions can occur for the particular case. The *true* causal effect for any single case cannot be estimated. Empirical inference of causal effects instead approximates the true causal effect. This is done by having several independent occurrences of the relation we are investigating. We then expose half of the cases to the condition X=1. This is the intervention group. The other half of the cases are kept the same without the intervention (X=0). For each case i exposed to the intervention, we measure the outcome $Y_i(X_i=1)$, and for each instance j not exposed to the intervention, we measure $Y_j(X_j=0)$ (Table 1). The mean difference between the two conditions estimates the true causal effect. The mean difference between the columns $Y(X=1)$ and $Y(X=0)$ in Table 1 is the *estimated* causal effect of X on Y (Rubin, 1974).

Table 1: Data frame for estimating the causal effect of variable *X* on outcome variable *Y*.

case	*X*	*Y*(*X*=1)	*Y*(*X*=0)	*Y*(*X*=1)-*Y*(*X*=0)
i_1	**1**	10	*NA*	*NA*
j_1	**0**	*NA*	0	*NA*
i_2	**1**	11	*NA*	*NA*
j_2	**0**	*NA*	0	*NA*
i_3	**1**	*9*	*NA*	*NA*
j_3	**0**	*NA*	0	*NA*
MEAN		**10**	**0**	**10**

For the causal effect to be a measure of causality—not just a correlation between two measurements—the intervention on *X* must be the only systematic change between conditions (Rubin, 1974; Woodward, 2012). In reality, however, there is always variation between single cases. The variation can come from several sources: imprecision in the measurements, noise in the environment, or variation inherent in what we denote as events of type *X*. We have to take variability between cases into account in the estimation of causal effects.

Random variation, unrelated to the intervention, is a minor problem as it will cancel out (to some degree) with an adequate number of cases. But if the variability between cases covaries with the intervention, it will invalidate the causal inference. It must be assumed that the intervention on *X* is the only variable that affects *Y* to make a valid causal inference. If systematic variation between groups occurs, it cannot be ruled out that the change in *Y* is due to confounding variables rather than *X*. It is important to control for systematic confounding *background variables* (i.e., any other variable than *X* and *Y*). Control of background variables is done by random sampling and systematic matching of background variables before making the intervention (Ahern et al., 2009; Rubin, 1974; Stuart, 2010). We can measure the background variables as separate variables B_1, B_2...B_n to ensure their distributions are similar between conditions. E.g., B_1 and B_3 in Table 2 appear to have similar distributions between cases *i* and *j*, but there seems to be a problematic difference between groups in B_2. Good experimental design requires adequate control of possible confounding variables.

Table 2: Expanded data frame for estimating the causal effect of *X* on *Y*.

case	B_1	B_2	B_3	*X*	*Y*(*X*=1)	*Y*(*X*=0)	*Y*(*X*=1)-*Y*(*X*=0)
i_1	5	1	3	**1**	10	*NA*	*NA*
j_1	4	10	2	**0**	*NA*	0	*NA*
i_2	6	0	1	**1**	11	*NA*	*NA*
j_2	5	11	3	**0**	*NA*	0	*NA*
i_3	4	2	2	**1**	*9*	*NA*	*NA*
j_3	6	9	1	**0**	*NA*	0	*NA*
MEAN	*i=5, j=5*	*I=1, j=10*	*i=1, j=1*		**10**	**0**	**10**

The estimated causal effect does not tell exactly what the true causal effect is for single cases: it is a generalized effect of type *X* events on type *Y* events (Dawid, 2000; Holland, 1986; Rubin, 1974). The estimated causal effect tells the probability of *Y* following *X* in the case that *Y* is a binary variable. If *Y* is a parametric variable, the causal effect is a numeric value indicating how much we expect *X* to change *Y* (Woodward and Hitchcock, 2003), e.g., ten "units" in the example in Table 1. In the following, whenever I refer to *causal effect*, I am referring to the *estimated causal effect*.

Causal effects are not *truths* in the logical sense. The causal effects are probabilistic relations between events estimated under controlled conditions that allow us to apply counterfactual logic to conclude a causal connection. For example, when testing if a new drug reduces fever, we do not need to describe how the chemical compound is absorbed in the body, passing the bloodstream, etc. to estimate the causal effect of the drug. As long as we have a causal effect of *X* on *Y* obtained under convincing circumstances, we can justify the conclusion that there is a causal relation between *X* and *Y*.

4 Causal inference for mental events

We can estimate the causal effects of mental events *M* in a similar way to how we estimate any other causal effect: the mental events are the variables we manipulate, and the behavioral outcome is the dependent variable we measure. Say we want to investigate if the intention to move one's arm cause one to move the arm: following the reasoning in the previous section, we can make an experimental manipulation so that subjects experience a specific mental event *M* (intention to move arm) in some conditions and not in others, while keeping all other variables constant, measure the outcome *Y* (physical movement of arm) for contrasting conditions, and estimate the causal effect of *M* on *Y*. *M* is a

cause of *Y* if there exists some intervention on *M* that changes the value of *Y* while keeping everything else equal (Woodward, 2012). The causal effect of *M* on *Y* is the mean difference between cases, for cases with *M*=1 and cases with and *M*=0, assuming there is no other systematic variation between the conditions. Causal inference with mental events is, like any other causal inference, estimated probabilistic relations between events. The inferred relations are the generalized effect of mental events of type *M* on type *Y* events.

To infer mental causation from experiments, we need several instances of the same mental event. We need to ensure that the mental events, which we conceptualize as experimental variable *M*, is similar enough across the entire experiment and between subjects that we can justify that they belong to the same type of mental event and estimate causal inference. E.g., if investigating the causal effect of the intention to move the hand, it must be assured that the intention is functional or phenomenological *equivalent* across cases. It is impossible to know if the *intention to move* is phenomenological identical between subjects as the experience is only available to the subjects, but this does not invalidate using it in causal inference. While we cannot assume that two similar behaviors are followed by similar mental states between subjects, we can make sure that the accompanying mental events are consistent. By measuring the mental events through introspective reports, we assess whether subjects describe the mental events in a consistent manner across subjects (Overgaard et al., 2008).

The problem of mental causation is often framed as whether *any* mental event can have *any* causal relevance in the physical world. This question is not suited for experimental research. Estimated causal effects apply to the events for which they were estimated. If we show one type of mental event M_a to be causally relevant (or irrelevant) for the outcome *Y*, then this does not mean that other mental events M_b, M_c, *etc*. have the same level of relevance. It is not given that the *experience of red* or the *intention to go on vacation* has the same causal relevance for moving one's arm as *the intention to move the arm* just because all are examples of mental events. They are different mental events. To conclude that the causal properties of one type of mental event apply to all mental events is an error analogous to concluding that the effect of one kind of drug applies to all kinds of drugs. Experimental studies of mental causation must be specific about what type of mental events they are dealing with.

In conclusion, to study mental causation with experimental research, we treat mental variables as any other experimental variable. But of course, if it is this simple, mental causation would not be a

controversial topic. The premise of this framework is that we do not know how mental events fit into the physical world. The unknown relation between mental events and physical events makes causal inference a peculiar enterprise. But it is not as difficult as one might think if we are aware of the mind-body problem.

4.1 Dealing with the special nature of mental events

To investigate the causal effects of mental events, we must first assume that mental events are related to the physical world. We do not need to assume *how* mental events are related to the physical world, we only assume that mental events *depend* on the physical world. This assumption is easy to justify: sensory inputs from the environment cause perceptions, intentions are directed towards the physical world, and particular neural activity is related to mental content (Aru et al., 2012; Chalmers, 2000).

We can make this assumption explicit: for each mental event *M*, there exists at least one physical event *P*, which is the minimal physical event necessary for instantiating *M* (for the given occurrence of *M*). This is different from stating that mental events are physical events. If we prefer, we can view mental events as non-physical properties. But if so, they are attached to the physical events that cause them.

How mental and physical events interact is still unknown: do the interaction go from the physical to the mental or can it go both ways? If the relation between *M* and *P* is one-directional, then mental events cannot produce causal effects. If mental events are causally irrelevant, then we hardly need an empirical approach to study them.

To outline the problem: we want to know if *M* can cause *Y* given *M* is bound by its physical base *P* (Figure 1). How we view the relation between *M* and *P* determines if this problem shows that mental causation is impossible.

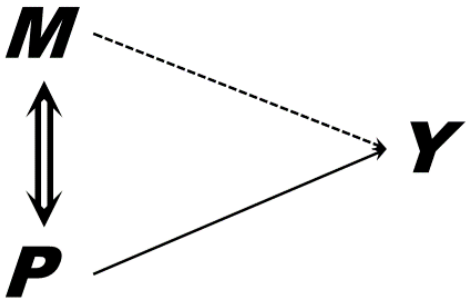

Figure 1: Outline of the problem: Can *M*—realized by *P*—be a cause of *Y*?

The first solution is to assume an *identity* relation between *M* and *P*, where the two events are the same singular event: any difference in appearance is only epistemological (e.g., Smart, 1959). The apparent dichotomy between physical and mental stems from the event being measured as either the physical variable *P* or as the mental variable *M*. The statements "*M* causes *Y"* and "*P* causes *Y"* are describing the same relation. If *P* is the cause of *Y* then so is *M* by definition. In the case of *M* and *P* sharing an identity, there is one causal factor that can influence *Y*.

In this case, mental variables are another type of physical variable, and it is valid to treat them like any other variable for the purpose of causal inference. If we measure either *M* as a mental variable or its physical base *P*, we would measure the same event twice. If we observe only one of either *M* or *P*, we automatically have proof that the other identity is present as well. This also gives that intervening on either *M* or *P* is an intervention on the same event. For inferential purposes, we must collapse *M* and *P* into a single causal factor *MP* (Figure 2).

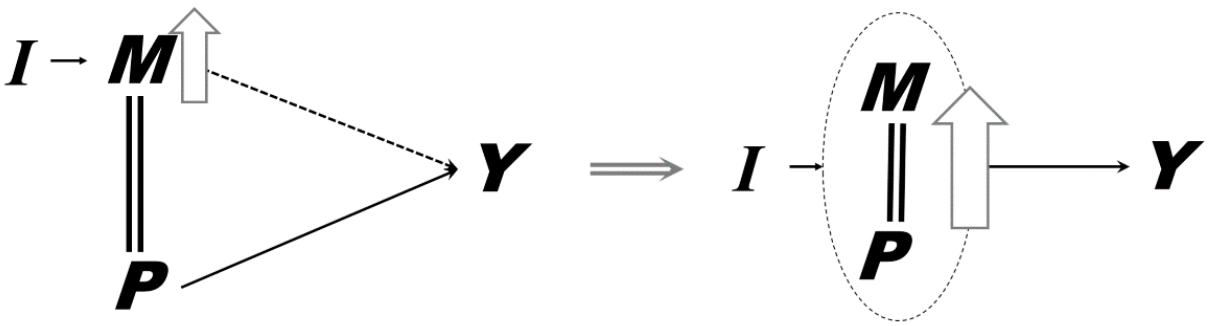

Figure 2: In the case of *M=P*, any intervention *I* that change *M* will be an intervention on *MP*.

On the other hand, we can take *M* to be different from *P*. The occurrence of *M* is still dependent on *P* (given the initial assumption), but *M* is a (non-reductive) supervenient property of *P*. The difference in appearance is not just epistemological: the difference between a mental variable measuring *M* and a physical variable measuring *P* each captures some underlying features that are non-overlapping.

Though *M* and *P* are ontologically distinct, *P* is both a sufficient and necessary condition for *M*. For *M* to be present so must *P* and when *P* is present, so is *M*. For *M* to be the cause of *Y*, *P* must be a physical cause of *Y*. To avoid *Y* being overdetermined there can only be one cause of *Y*, so either *M* or *P* must be removed as the cause of *Y*. If the physical world is causally closed, then *P* cannot be eliminated. Hence, *P* is a sufficient cause of any change in *Y* we would ascribe to *M*. *M* is an epiphenomenon and is excluded from the causal relation (Kim, 2005). Only the solid arrow in Figure 1 describes the real causal connection between *M*, *P*, and *Y*.

If we take *M* as a supervenient property of *P*, then mental variables measure different phenomena than the physical variables do. But since *P* is sufficient and necessary for *M*, we run into problems if we try to isolate either and estimate causal effects.

Imagine an experiment where we have a brain stimulator that can target—and only targets—non-physical mental events without affecting any physical events. We use the non-physical stimulator to induce the intention to move one's arm (M=1) in a group of subjects. We also have a control group that is not subjected to the non-physical stimulation and will not experience the intention to move their arms (M=0). Assume that the brains of the subjects, independent of group, all are in a given state P^* at the moment before the intervention. P^* is in no way related to the intention to move one's arm. During the experiment, the brains of the control group will continue to be in state P^*. When the non-physical stimulator induces the (non-reductive) intention to move in the intervention group, it must follow that the physical base *P* of the intention to move have to be present for *M* to be present. For the non-physical intervention to change the value of *M*, it will follow that P^* change to *P* (Baumgartner, 2009; Kim, 2005). It is impossible for *M* to change without a corresponding change in *P*.

That both *M* and *P* change is a problem for causal inference: since $P \neq M$, we must place *P* as a column in our matrix of confounding variables (Table 3). But if *P* is not present then neither is *M*: *M* and *P* covary. We can estimate a difference between the two groups, but we cannot determine if the effect is

caused by *M* or *P*: it is not possible to isolate the non-physical *M* as required for causal inference (Baumgartner, 2010, 2009). We cannot estimate the causal effect of a mental event alone.

Table 3: Data frame for estimating the causal effect of mental event *M* on *Y*. *M* and its physical base *P* are perfect covariates.

case	B_1	B_2	*P*	*M*	*Y*(*X*=1)	*Y*(*X*=0)	*Y*(*X*=1)-*Y*(*X*=0)
i_1	5	3	**1**	**1**	10	*NA*	*NA*
j_1	4	2	**0**	**0**	*NA*	0	*NA*
i_2	6	1	**1**	**1**	11	*NA*	*NA*
j_2	5	3	**0**	**0**	*NA*	0	*NA*
i_3	4	2	**1**	**1**	*9*	*NA*	*NA*
j_3	6	1	**0**	**0**	*NA*	0	*NA*
MEAN					**10**	**0**	**10**

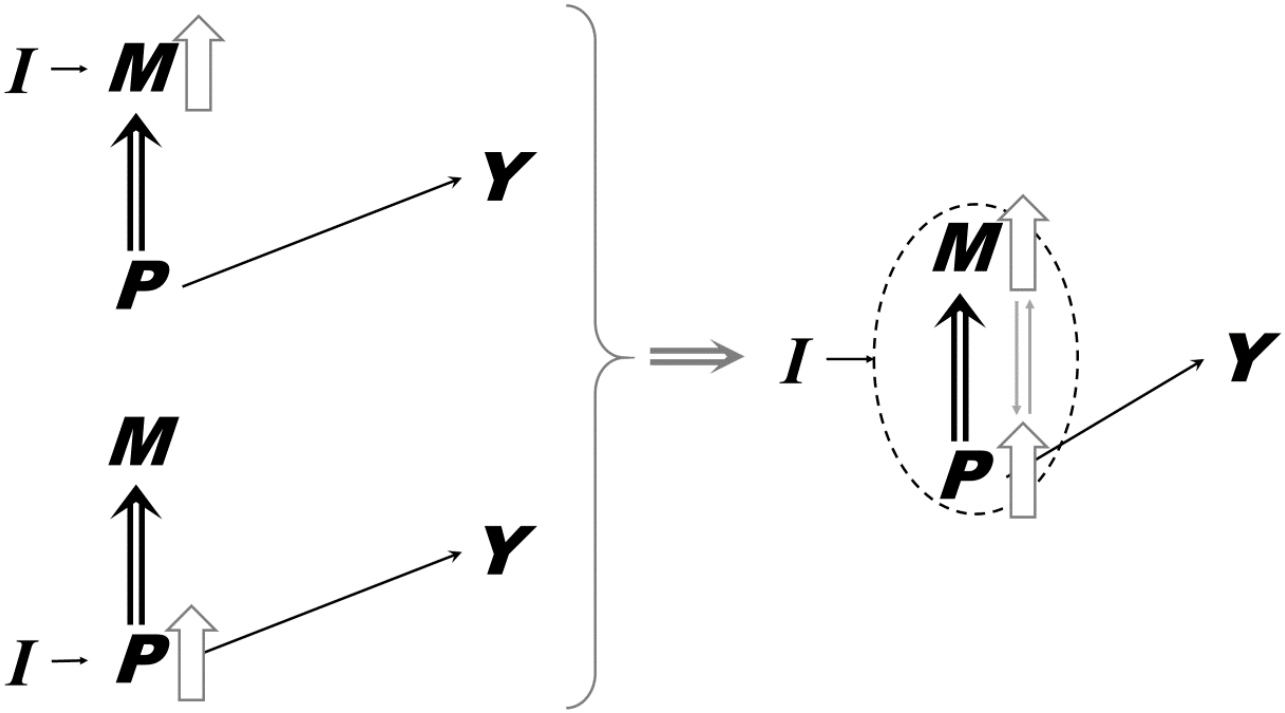

Figure 3: In case *M* is a (non-reductive) supervenient property of *P*, any intervention *I* on *M* must be accompanied by a corresponding change in *P*. *M* and *P* cannot be separated as causal factors despite being ontologically different. The

same is true if we make an intervention *I* on *P*: Any intervention on either *M* or *P* will be an intervention on *MP* as a change in one will be accompanied by a change in the other.

The covariation between *M* and *P* is not only an issue when isolating *M* as a causal factor. Imagine the same experiment above, but this time we use a different brain stimulator that intervenes on *P* and estimates the causal effect of *P* on *Y* (Figure 3, bottom diagram). In the intervention group, the brain-states change from *P** to *P* while the brain-states of the control group remain the same. Since *P* is the physical base of *M* then a change from *P** to *P* will also induce *M*. As with the non-physical stimulation inducing *M*, we are not able to isolate *P* as the cause of *Y*. Even if we assume that *M* and *P* are different, it is impossible to separate them as causal factors: *M* and *P* are perfect covariates.

Since we cannot separate *M* and *P* as causal factors, we have two options when it comes to causal inference: we can abandon the attempt of an empirical approach to mental causation, as we cannot construct an experiment where we isolate *M* and keep all other factors (including *P*) constant. Or we can collapse *M* and *P* into a single variable *MP*.

Considering causal properties of mental events and their physical base as a single causal factor is not new in the analytical approach to mental causation (e.g., Kim, 2005; Lewis, 1994; Mele, 2009; Sperry, 1980; Woodward, 2015). Here I argued for the same position as we cannot separate the two in practice: it is a practical necessity—not an ontological assumption. *M* and *P* can be ontologically different, but the isolation of the factor *M* from *P* is impossible no matter which of the above solutions to the mind-body problem we prefer. This means that we cannot answer the question if a mental event or its physical base is the cause of action by empirical means. We can only answer if the mental event and its physical base as a single factor is the cause of action. For practical purposes, the causal relevance of a mental event *M* is the same as its physical base *P*.

4.2 Reduction and causal explanations

One could argue that we are reducing away *real* mental causation by always considering the causal properties of mental events together with their physical realization. This is only the case if we start with the position that mental causation per definition must be non-physical. If one feels that the concept of mental causation is reserved for "pure" mental causation, we can instead call the causal effects "mental-and-physical-base causation" and proceed.

One can still ask: since the causal effect of *M* only can be estimated together with *P,* can we then remove *M* from the causal relation and only deal with the physical variables? In principle, we can remove *M* from the causal explanation. But it is not feasible in practice. In reality, it is rarely the case that higher-order explanations can be reduced to statements involving fundamental physical processes, even in cases where we are justified in assuming the higher order phenomena are genuine reducible physical phenomena (Anderson, 1972; Bedau, 2002). The reason is that we do not understand what constitutes a physical base of a mental event. What real type of physical event *P* denotes in the examples above is undisclosed in the real world. It might be explanations involving neural anatomy, neural communication, or fundamental laws of physics. Even if we assume mental events are reducible, we do not have the relevant information to remove mental events from the inference of mental causation.

Reductionism in practice is also problematic in the experimental setting. To estimate causal effects, we need several instances of the same mental event *M,* and we cannot guarantee that each occurrence of *M* is identical. Though each case has to be similar enough for the causal inference to be valid, we have to tolerate some variation in practice. The variation within an experimental variable opens the possibility that each case of *M* does not have the same physical base. This can be because of the uncertainties in the mental variables or because mental events can be realized by different physical events (Fodor, 1974). Rather than thinking about *M* and *P* as ideal events, think of them as sets of similar events: $\boldsymbol{M}$ is the set of similar mental events $M_1, M_2 \ldots M_N$ and $\boldsymbol{P}$ is a set of physical events $P_1, P_2 \ldots P_N$ that each corresponds to the physical base of the mental events in $\boldsymbol{M}$. Since is possible for the variable *P* to be a set of several different physical states we cannot assume that all elements in $\boldsymbol{M}$ can be collapsed with the same physical event P_i. Each mental event M_i must have its own corresponding physical base P_i. When considering the causal properties of each M_i, we cannot distinguish it from the causal properties of its physical base P_i, and we have to collapse each M_i and P_i into a single causal factor. The problem of separating M and *P* does not change, but it does change how we can control for confounding physical variables in causal inference. We have to deal with a practical form of multiple realizations of events, even if the events are reducible (Aizawa & Gillett, 2009). It is, thus, problematic to attempt to completely remove mental variables, by only measuring physical variables, as we cannot be sure that

we measure the correct underlying physical base in ever repetition—at least not with our current understanding of how the brain gives rise to mental events.

Finally, removing mental events from inquiry about mental causation, even in ontologically possible, might not be desirable in practice. Low-level explanations do not always provide useful causal explanations. Causal explanations are about how some units of interest *X* have an impact on another unit of observation *Y*. The units of interest can be any phenomenon, e.g., brain processes, pills, social factors, or mental events. We can assume that *X* can be broken into pieces and explained as composites x_1, x_2, etc., and explain how the composites cause outcome *Y*. Replacing causal explanation involving *X* with an explanation about the composites have to take all the composite parts into account. The explanation involving the composites increase in complexity. The increase in complexity makes the required explanation more difficult and is contrary to the purpose of causal explanations (Lipton, 2005).

To say that a pill cause fever reduction does (in most cases) provide the information we want to know about the pill and its use—even if we have a full account of the chemical compounds in the pill and how they interact with the biophysical processes that regulate body temperature. Similar, to say that my intention to raise my arm is the cause raising my arm, is as valid a causal explanation as an explanation involving all the neural processes giving rise to my intention to raise my arm. One is not more correct than the other, but the first is a lot simpler than the latter. Since causal inference is relative effects of variables of interest, we can reverse the issue of reductionism and ask if we can remove physical variables and only consider the causal effects of mental variables? The proper level of analysis depends on the phenomena we are interested in. It is not necessary to regress to lower level explanations to explain the causation of higher-order events to address causation.

The reason we want physical variables is that in the context of mental causation this is usually part of what we want to know: it is not if mental events can have causal effects, it is *how* they can have causal effects and how they interact with physical processes; e.g., how the intention to move the hand is part of the nervous system responsible for locomotion. How does the relative contribution from mental events and non-mental physical processes generate behavior? By combining mental variables and (physical) neuro-cognitive variables in experimental designs that we can answer these questions.

5 Towards a science of mental causation

Combining mental variables and physical variables in causal inference is difficult, as we are dealing with variables who's underlying ontology can be dependent on one another. When measuring blood flow in the brain or electrophysiological potentials, we do not know whether the measured activity is the physical base of mental event *M*, unrelated to *M*, or only a part of the physical base. The mental event and its physical base *MP* is not the only event occurring in the brain when we investigate the causal effect of *MP*. Isolating the neural base of a mental event is not as simple as observing which physical variables that co-vary with the mental variables. Any physical variable we measure can be a precursor of the real physical base of the mental event or an effect following the mental event (Aru et al., 2012). Since we are dealing with complex systems and different explanatory levels, we have to be careful when operationalizing potential causal factors in experiments.

To study causal processes in cognitive science is difficult, even without considering mental variables. The nervous system is not wired as a linear causal chain where *X* causes *Y*, causes *Z*, and so on. The brain consists of interconnected networks that operate on different anatomical scales and different timescales. For example, to investigate causal effects of *the intention to move the hand* on hand movements, we have to consider that the process is part of a network that depends on both long-range connectivity and local specification of function working in many hierarchical feedback loops (Rizzolatti and Luppino, 2001; Shadmehr and Krakauer, 2008).

To ask whether a mental event *M* cause action *Y* is not different from asking if certain neural activity in the supplementary motor area (call this activity P_{SMA}) cause *Y*. To infer this relation following the principles of causal inference, we make an intervention on P_{SMA} while keeping everything else in the brain is unaffected. If we find that the intervention on P_{SMA} changes *Y*, compared to an adequate control condition, we can conclude that P_{SMA} causes *Y*. This does not mean P_{SMA} is the only cause of *Y* nor that P_{SMA} is an isolated cause of *Y*. We should not fool ourselves to believe that because we measure two events (P_{SMA} and *Y*) in a complex system (the brain and its interaction with the environment) and study the effect, then nothing else of relevance is going on. There definitely would be in this example, e.g., the neural communication from cortical motor areas to the basal ganglia, thalamus, and the peripheral

nervous system, but in the experiment, we make sure that all these are constant to pursue the question aimed at P_{SMA}.

We cannot make inference about all the other variables from the experiment, we can only make an inference about the causal effect of P_{SMA}. To ask if a single variable X causes Y is a simplification of the complex mechanisms we are dealing with. But it is a simplification for inferential purposes, not a simplification of how the system is. Estimating a causal effect of X on Y does not tell us about the possible effects of other events Z_1, Z_2, etc. on Y or how they are related. It only gives the causal effect of X on Y. To test whether Z_1, Z_2, etc. are causes of Y, we need separate experiments for each type of event. The "start" and "end" of the relation are pragmatic cuts where we select events of the kind that we want to investigate from the larger configuration.

Controlling variables "downwards" become increasingly difficult. Since we do not know the precise relation between the mental and physical, we might try to control for a physical variable that, unknown to us, is a part of the physical base of the mental event.

The physical base of mental events is likely a complex system of interconnected processing between distinct sub-parts in a network (Baars, 2005; Tononi and Koch, 2015). Each node in the network is not enough to constitute the physical base of M. Only all parts connected are sufficient to enable M.

Let us continue the example above to make inference about the lower level configuration of P_{SMA}. SMA is divided into fine-grained anatomy based on local functionality and afferent connections (Nachev et al., 2008). Assume we can divide P_{SMA} into four parts, as illustrated in Figure 4. All parts have to be "active" to constitute P_{SMA} (and thereby M). By replacing the unified base P with the parts in the data-frame for the causal experiment, we get Table 4. All parts, which together constitute the base of M, are perfect covariates with M as P were in Table 2.

When we look at the parts, only P'_4 has a direct causal link to outcome Y. If we were to intervene on P'_4 in Figure 4 (keeping all other variables constant), we find that P'_4 have a causal effect on Y. This time M is no longer a perfect co-variate: we see that M only occurs when all sub-components are present, as shown in Table 5. The conclusion we would draw from Table 4 (M cause Y) and Table 5 (P'_4 cause Y) are both correct.

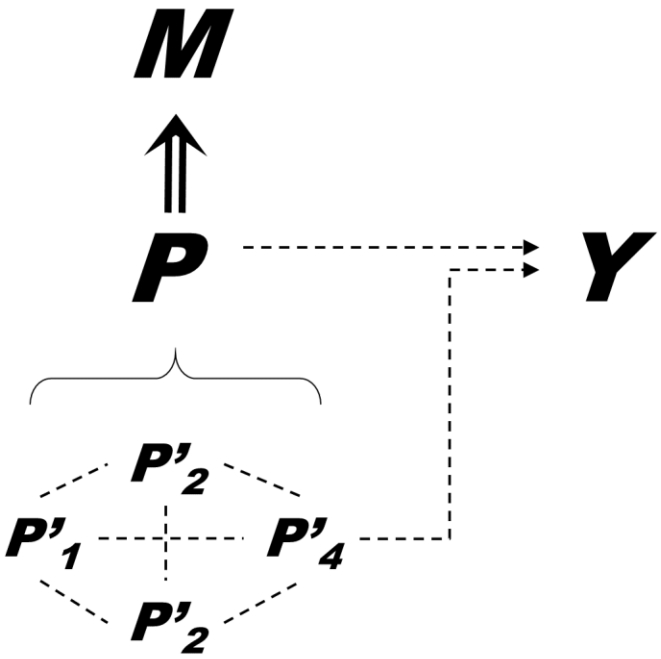

Figure 4: M is realized by P, which is constituted by the parts P'$_1$, P'$_2$, P'$_3$, and P'$_4$. Om sub-part P'$_4$ has a connection to Y.

Table 4: Data frame for estimating the causal effect of mental event *M* on outcome variable *Y* with measurements of low-level parts of *P*.

case	B_1	B_2	P'_1	P'_2	P'_3	P'_4	*M*	*Y*(*X*=1)	*Y*(*X*=0)	*Y*(*X*=1)-*Y*(*X*=0)
i_1	5	3	**1**	**1**	**1**	**1**	**1**	10	*NA*	*NA*
j_1	4	2	**0**	**0**	**0**	**0**	**0**	*NA*	0	*NA*
i_2	6	1	**1**	**1**	**1**	**1**	**1**	11	*NA*	*NA*
j_2	5	3	**0**	**0**	**0**	**0**	**0**	*NA*	0	*NA*
i_3	4	2	**1**	**1**	**1**	**1**	**1**	*9*	*NA*	*NA*
j_3	6	1	**0**	**0**	**0**	**0**	**0**	*NA*	0	*NA*
MEAN								**10**	**0**	**10**

Table 5: Data frame for estimating causal effect similar to Table 4, but for the causal effect of a low-level part of *P* on outcome variable *Y* instead of mental event *M*.

case	B_1	B_2	P'_1	P'_2	P'_3	M	P'_4	$Y(X=1)$	$Y(X=0)$	$Y(X=1)$-$Y(X=0)$
i_1	5	3	**1**	**1**	**1**	**1**	**1**	10	*NA*	*NA*
j_1	4	2	1	0	1	0	**0**	*NA*	0	*NA*
i_2	6	1	0	1	0	0	**1**	11	*NA*	*NA*
j_2	5	3	1	1	0	0	**0**	*NA*	0	*NA*
i_3	4	2	1	0	0	**0**	**1**	*9*	*NA*	*NA*
j_3	6	1	0	0	1	0	**0**	*NA*	0	*NA*
MEAN								**10**	**0**	**10**

It might appear paradoxical that *M* (together with *P*) can both be a cause of *Y* and, at the same time, be controlled when estimating the causal effect of only a part of *P*. The apparent discrepancy is because the two tables represent two different causal experiments. Causal effects are about the variable manipulated by the intervention. They do not tell us about the causal impact of other factors. We can for example not conclude that there is no effect of P'_1, P'_2, P'_3, or *M* from the experiment in Table 5. We can only conclude that there is an effect of P'_4 om *Y*. How we define the events, we investigate, determines the explanations we can give based on the causal inference.

To conclude that "higher order" events are causally irrelevant is not possible from the observation that only a subset of low-level components has a causal effect on the outcome. Demonstrating that P'_4 is the part of *P* that cause *Y*, does not mean that *P* is not also a cause of *Y*. The explanation containing P'_4 is a fine-grained causal explanation that the one containing *P* or *M*, but it does not follow the higher order elements are irrelevant. Stating that a mental event is the cause of action does not exclude causal explanations in terms of neural mechanisms, nor does causal explanations in terms of neural mechanisms exclude causal explanations involving mental events (Pernu, 2011).

6 Inferring mental and non-mental causes

It is tempting to treat mental variables and physical variables as measures of different ontological levels and contrast the two variables to answer whether the physical event or mental event is the cause of *Y*. But this is not possible. Mental events are realized by physical events that might be contained in the measured physical variable. Similar, we cannot per default claim that no mental events are occurring when we measure physical variables; especially when the type of measured physical event is part of the organ that generates the mental events. It is only possible to claim that a physical variable represents an

unconscious (non-mental) cause of action if no mental variables influence the outcome. It is not enough to show that a physical variable influences the outcome to conclude that mental events are irrelevant for action. For example, movement-related cortical potentials precede the intention to move the hand (Fried et al., 2011; Libet et al., 1983), which is taken to prove that the intention to move is not a cause of moving the hand—the true cause is unconscious neural processes expressed as the movement-related cortical potentials (Harris, 2012; Libet, 1999, 1985; Wegner, 2002).

Showing that a physical variable (movement-related potentials) precedes a mental variable (the intention to move) is not enough to infer that the former is the cause of outcome rather than the latter. The observation lacks a contrasting condition to rule out if there is a causal effect of the intention on the outcome. It is impossible to infer that the intention is *not* a cause of moving the hand from the observation. It is even impossible to infer that the movement-related potentials are the cause of moving the hand. Precedence does not imply causation. We can only claim that an action is unconsciously initiated if we show both a causal effect of the unconscious processes and a null-effect of the mental event. To infer true unconscious causes of action, we must include mental events as background variables to show that mental events do not covary with the unconscious events.

For example, it is unclear how the movement-related potentials mentioned above are related to the conscious intention to move. Whether they reflect unconscious neural activity is unknown. Some studies have shown covariation of the readiness potential and conscious intention, while others are unable to do so (Haggard and Eimer, 1999; Keller and Heckhausen, 1990; Schlegel et al., 2013; Schultze-Kraft et al., 2016; Vinding et al., 2014). To conclude that a physical variable represents genuine unconscious action initiation it must be shown that it has a causal effect while keeping the mental content constant.

It is surprisingly difficult to determine what *non-mental* or unconscious means (Moors and De Houwer, 2006). It is not as simple as dividing mental/conscious and non-mental/non-conscious processes: the transition can be gradual (Miller and Schwarz, 2014; Sandberg et al., 2011), and there are separate ways to be unconscious of stimuli (Kim and Blake, 2005; Rothkirch and Hesselmann, 2017). In conclusion: it is not valid to ignore mental events altogether and conclude that the action was unconsciously initiated. If we want to show that a low-level neural event is the cause of action rather than mental events, then it requires a null-effect of the relevant mental variable.

7 Conclusion

The complexity of the nervous system makes investigating mental causation a difficult task. But it is possible to study mental causation by applying the principles of causal inference as in any scientific field. Given the special nature of mental events, we need to treat the mental event and its physical base as one factor in the experimental design. This means that we cannot answer whether a mental event or its physical base is the cause of action. The type of questions we can answer is whether a given mental event *M* (realized by *P*) is a cause of *Y*. Mental causation is measured as the causal effect of *MP* on *Y* in controlled experiments with contrasting conditions that control for confounding variables. This approach is not for those who seek an either/or answer to mental causation, but it is of relevance to those who seek to investigate the neurocognitive and behavioral relevance of mental events. Causal effects are in any instance relative contributions of the variables—not ontological truths.

To investigate the mental causation, we have to think about mental causation as relative contributions of events in complex systems with different descriptive levels. The different descriptive levels do not preclude one another. Experimental scientists must shift their approach to mental causation from the search for ultimate answers in the analytical discussion and instead focus on relative effects of well-operationalized variables. Cognitive scientists have to consider how to manipulate mental events in experimental design and how to control confounding variables. The solution will depend on the type of mental event and outcome behavior in question. Note that these considerations are methodological, not metaphysical.

Acknowledgments

This work was funded by Danish Independent Research Council | Humanities. I would like to thank Lise Marie Andersen and Nina Jensen for comments on earlier drafts of this paper.

References

Ahern, J., Hubbard, A., Galea, S., 2009. Estimating the Effects of Potential Public Health Interventions on Population Disease Burden: A Step-by-Step Illustration of Causal Inference Methods. Am. J. Epidemiol. 169, 1140–1147.

Anderson, P.W., 1972. More Is Different. Science 177, 393–396.

Aru, J., Bachmann, T., Singer, W., Melloni, L., 2012. Distilling the neural correlates of consciousness. Neurosci. Biobehav. Rev. 36, 737–746.

Baars, B.J., 2005. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In: Progress in Brain Research. Elsevier, pp. 45–53.

Baumgartner, M., 2009. Interventionist Causal Exclusion and Non-reductive Physicalism. Int. Stud. Philos. Sci. 23, 161–178.

Baumgartner, M., 2010. Interventionism and Epiphenomenalism. Can. J. Philos. 40, 359–383.

Bedau, M., 2002. Downward Causation and the Autonomy of Weak Emergence. Principia 6, 5–50.

Block, N., 2005. Two neural correlates of consciousness. Trends Cogn. Sci. 9, 46–52.

Chalmers, D.J., 1997. The conscious mind. Oxford University Press, New York; Oxford.

Chalmers, D.J., 2000. What is a neural correlate of consciousness? In: Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press, pp. 17–40.

Cohen, M.A., Dennett, D.C., 2011. Consciousness cannot be separated from function. Trends Cogn. Sci. 15, 358–364.

Crane, T., Mellor, D.H., 1990. There is no question of physicalism. Mind 99, 185–206.

Dawid, A.P., 2000. Causal Inference without Counterfactuals. J. Am. Stat. Assoc. 95, 407–424.

Fodor, J.A., 1974. Special sciences (or: the disunity of science as a working hypothesis). Synthese 28, 97–115.

Fried, I., Mukamel, R., Kreiman, G., 2011. Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition. Neuron 69, 548–562.

Haggard, P., Eimer, M., 1999. On the relation between brain potentials and the awareness of voluntary movements. Exp. Brain Res. 126, 128–133.

Harris, S., 2012. Free will. Free Press, New York.

Hohwy, J., 2009. The neural correlates of consciousness: New experimental approaches needed? Conscious. Cogn. 18, 428–438.

Holland, P.W., 1986. Statistics and Causal Inference. J. Am. Stat. Assoc. 81, 945–960.

Jensen, M., Dong, M., Vinding, M.C., Overgaard, M., 2017. Measuring sensation of movement. In: Grünbaum, T., Christensen, M.S. (Eds.), Sensation of Movement, Current Issues in Consciousness Research. Routledge, New York, NY.

Keller, I., Heckhausen, H., 1990. Readiness potentials preceding spontaneous motor acts: voluntary vs. involuntary control. Electroencephalogr. Clin. Neurophysiol. 76, 351–361.

Kim, C.-Y., Blake, R., 2005. Psychophysical magic: rendering the visible 'invisible.' Trends Cogn. Sci. 9, 381–388.

Kim, J., 2005. Physicalism, or something near enough, Princeton monographs in philosophy. Princeton University Press, Princeton, N.J.

Laureys, S., 2005. The neural correlate of (un)awareness: lessons from the vegetative state. Trends Cogn. Sci. 9, 556–559.

Lewis, D., 1994. Reduction of Mind. In: A Companion to Philosophy of Mind. Blackwell Publishers, Oxford, pp. 412–431.
Libet, B., 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. Behav. Brain Sci. 8, 529–566.
Libet, B., 1999. Do we have free will? J. Conscious. Stud. 6, 47–57.
Libet, B., Wright, E.W., Gleason, C.A., 1983. Preparation-or intention-to-act, in relation to pre-event potentials recorded at the vertex. Electroencephalogr. Clin. Neurophysiol. 56, 367–372.
Lipton, P., 2005. Inference to the best explanation, 2nd ed. ed, International library of philosophy. Routledge, London.
Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.
Mele, A.R., 2009. Effective intentions: the power of conscious will. Oxford University Press, Oxford ; New York.
Melnyk, A., 1997. How to Keep the "Physical" in Physicalism. J. Philos. 94, 622.
Miller, J., Schwarz, W., 2014. Brain signals do not demonstrate unconscious decision making: An interpretation based on graded conscious awareness. Conscious. Cogn. 24, 12–21.
Moors, A., De Houwer, J., 2006. Automaticity: A Theoretical and Conceptual Analysis. Psychol. Bull. 132, 297–326.
Nachev, P., Kennard, C., Husain, M., 2008. Functional role of the supplementary and pre-supplementary motor areas. Nat. Rev. Neurosci. 9, 856–869.
Overgaard, M., Gallagher, S., Ramsøy, T.Z., 2008. An integration of first-person methodologies in cognitive science. J. Conscious. Stud. 15, 100–120.
Pernu, T.K., 2011. Minding matter: how not to argue for the causal efficacy of the mental. Rev. Neurosci. 22.
Reingold, E.M., Merikle, P.M., 1988. Using direct and indirect measures to study perception without awareness. Percept. Psychophys. 44, 563–575.
Rizzolatti, G., Luppino, G., 2001. The Cortical Motor System. Neuron 31, 889–901.
Rothkirch, M., Hesselmann, G., 2017. What We Talk about When We Talk about Unconscious Processing – A Plea for Best Practices. Front. Psychol. 8.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.
Sandberg, K., Bibby, B.M., Timmermans, B., Cleeremans, A., Overgaard, M., 2011. Measuring consciousness: Task accuracy and awareness as sigmoid functions of stimulus duration. Conscious. Cogn. 20, 1659–1675.
Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P.U., Wheatley, T., 2013. Barking up the wrong free: readiness potentials reflect processes independent of conscious will. Exp. Brain Res.
Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., Blankertz, B., Haynes, J.-D., 2016. The point of no return in vetoing self-initiated movements. Proc. Natl. Acad. Sci. 113, 1080–1085.
Seth, A.K., Baars, B.J., Edelman, D.B., 2005. Criteria for consciousness in humans and other mammals. Conscious. Cogn. 14, 119–139.
Shadmehr, R., Krakauer, J.W., 2008. A computational neuroanatomy for motor control. Exp. Brain Res. 185, 359–381.
Smart, J.J.C., 1959. Sensations and Brain Processes. Philos. Rev. 68, 141.

Smart, J.J.C., 1978. The Content of Physicalism. Philos. Q. 28, 339.
Sperry, R.W., 1980. Mind-brain interaction: Mentalism, yes; dualism, no. Neuroscience 5, 195–206.
Stoljar, D., 2001. Two Conceptions of the Physical. Philos. Phenomenol. Res. 62, 253.
Stuart, E.A., 2010. Matching Methods for Causal Inference: A Review and a Look Forward. Stat. Sci. 25, 1–21.
Tononi, G., Koch, C., 2015. Consciousness: here, there and everywhere? Philos. Trans. R. Soc. B Biol. Sci. 370, 20140167–20140167.
Vinding, M.C., Jensen, M., Overgaard, M., 2014. Distinct electrophysiological potentials for intention in action and prior intention for action. Cortex 50, 86–99.
Wegner, D.M., 2002. The illusion of conscious will. MIT Press, Cambridge, Mass.
Woodward, J., 2005. Making things happen: a theory of causal explanation, Oxford studies in philosophy of science. Oxford University Press, Oxford.
Woodward, J., 2012. Causation: Interactions between Philosophical Theories and Psychological Research. Philos. Sci. 79, 961–972.
Woodward, J., 2015. Interventionism and Causal Exclusion. Philos. Phenomenol. Res. 91, 303–347.
Woodward, J., Hitchcock, C., 2003. Explanatory Generalizations, Part I: A Counterfactual Account. Nous 37, 1–24.