Synthese (2018) 195:5159–5174
https://doi.org/10.1007/s11229-018-1713-z

S.I.: NEUROSCIENCE AND ITS PHILOSOPHY

Mental causation via neuroprosthetics? A critical analysis

Tuomas K. Pernu[1,2]

Received: 1 June 2017 / Accepted: 31 January 2018 / Published online: 21 February 2018

Abstract Some recent arguments defending the genuine causal efficacy of the mental have been relying on empirical research on neuroprosthetics. This essay presents a critical analysis of these arguments. The problem of mental causation, and the basic idea and results of neuroprosthetics are reviewed. It is shown how appealing to the research on neuroprosthetics can be interpreted to give support to the idea of mental causation. However, it does so only in a rather deflationary sense: by holding the mental identical with the neural. So contrary to what the arguments have been assuming, neuroprosthetics cannot be used to argue for nonreductive physicalism. It can rather be taken to illustrate just the opposite: how the mental and the physical are identical.

Keywords Action control · Brain-computer interface · Causal exclusion · Difference-making · Interventionism · Multiple realisability

1 Introduction

Imagine the following scenario. You wake up in a bed, drowsy and tired, not able to locate yourself. A nurse steps into the room, walks by your bed and gives you a briefing: you have suffered a serious stroke some days earlier, you are in a hospital, and you are coming out of a coma. Everything is confusing.

✉ Tuomas K. Pernu
tuomas.pernu@kcl.ac.uk

1 Department of Philosophy, King's College London, Philosophy Building, Strand, London WC2R 2LS, UK

2 Molecular and Integrative Biosciences Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

You lie in the bed trying to put all the pieces together. Soon you realise that you can't feel your limbs. You try to move them, but nothing is happening. A doctor walks into the room and comes to your bed. She asks how you feel and starts to examine you. You respond, with a feeble voice: "I cannot feel my limbs". The doctor looks at you and calmly explains that it seems, as they had expected, that you have suffered a serious paralysis, and it is most likely that you will never be able to move your limbs again. Before you can really understand what you have just been told she continues reassuringly: you are lucky; the nature of your condition is such that there are new devices, special prosthetics, that can help you to interact with your environment. These devices, you are being explained, could be connected directly to your nervous system, and you would be able to control them, after a little practice, as fluently as your own limbs. In fact, pretty much immediately, they would actually start *feeling* like your own limbs. You find this all difficult to understand, and you are still very confused about everything, but you are also exhausted. You fall back asleep.

This is not science fiction. Neuroprosthetics is part of today's medicine. The field is developing rapidly, and procedures that connect artificial devices directly to the nervous system are becoming more commonplace and wide-spread. Roughly half a million people worldwide have a cochlear implant, an electronic hearing aid that connects to the cochlear nerve in the ear. Individual patients are being tested with robotic arms and other mechanical devices that function like normal limbs. The more we understand about the neural basis of cognition and motor coordination, and the better we become in biomedical engineering, the more diverse devices we will be able to develop and the more routinely we will be implanting them.

Neuroprosthetics are tremendously helpful; there is no question about their medical usefulness. But the procedures are invasive, and they raise—quite understandably—a number of burning ethical questions (e.g. Clausen 2008, 2011; Glannon 2016). Recently, however, the field has been receiving quite a different sort of philosophical attention: neuroprosthetics has been used to illustrate how the mind can have genuine causal power—that there can be such a thing as mental causation (e.g. List and Menzies 2009; Menzies 2015; Woodward 2008a, b, 2017). These arguments rely on an interventionist or a difference-making notion of causation, the proponents of which are eager to stress how this particular notion of causation is most appropriately aligned with the current scientific practice. It is thus not difficult to see how enticing indeed it may seem to appeal to some recent research in neuroprosthetics.

The following will present a critical analysis of these arguments. The problem of mental causation, the difference-making response to it, and the basic idea and results of neuroprosthetics are reviewed. It is shown how appealing to neuroprosthetics can be interpreted to give support to the idea of mental causation. However, it does so only in a rather deflationary sense: by holding the mental identical with the neural. In fact, neuroprosthetics illustrates quite vividly how our ability to identify the neural basis of mental states constitutes major advances in science—and how these advances can be turned into concretely helpful applications. So contrary to what the arguments have been assuming, neuroprosthetics cannot be used to argue for nonreductive physicalism. It can rather be taken to illustrate just the opposite: how the mental and the physical are identical.

The main lesson of the following analysis is that the difference-making argumentation has been equivocal in important respects: it has largely ignored the question of how, or indeed whether, the "mental" and the "physical" are distinct. In the argumentation neuroprosthetics is used to illustrate the idea how higher-level, coarse-grained causal hypotheses can, and often should be, preferred to lower-level, fine-grained causal hypotheses. However, even if one would accept this conclusion, to save nonreductive physicalism—and the idea of genuine, autonomous mental causation—one would need to show further that this sort of relationship holds between the mental and the physical. That is, one would need to show that mental states are in fact multiply realised. But having a closer look at neuroprosthetics actually results in an opposite verdict: neuroprosthetics works through identifying mental states with physical (*i.e.* cortical) states. Therefore, if one wants to hold on to neuroprosthetics as a paradigmatic example of mental causation, then one should give up nonreductive physicalism. Or, if one is not prepared to compromise nonreductive physicalism, then one should be addressing more explicitly the issue of how the mental and the physical should be construed as distinct, and give up the idea that neuroprosthetics could be used to back up such a distinction.

2 Causal exclusion and difference-making

The idea that we could have genuine, autonomous mental causation in an ultimately physical world faces the problem of causal exclusion (Kim 1998, 2005). Suppose that every mental state (M) is always realised by some neural state (N), as required by physicalism, and suppose that we are trying to account for a subsequent behaviour (B). Now physicalism (the causal completeness of the physical) requires also that N is causally sufficient for B, that N is all that we need to account for the occurrence of B. Now M, supposing that it is distinct from N, seems to be left with a thoroughly otiose role: for all our behaviour there are always prior neural states that are causally sufficient for the behaviour, and hence the mental states supervening on the neural states function only as overdeterminers. Since systematic overdetermination is not acceptable, we need to conclude that mental states are void of causal power.

The difference-making response to the causal exclusion problem proceeds in two steps. First, the extensive use of the notion of "causal sufficiency" in the causal exclusion argument is criticised: "[t]he reference to causal sufficiency harks back to older empiricist accounts of causation that take causation simply to be some form of subsumption under laws" (List and Menzies 2009, p. 490). It thus seems that the exclusion argument depends "on somewhat outmoded assumptions from deductive-nomological accounts of causation and causal explanation" (List and Menzies 2009, p. 490). Hence we should, according to these arguments, discard the notion of causal sufficiency, and only speak in terms of causation *simpliciter*.

In the next step, causation is defined in terms of difference-making. The idea, in a nutshell, is the following. For a property C to cause the presence of another property E (in a world w) the following pair of counterfactuals must hold (in w):

(a) $C \mathbin{\Box\!\!\rightarrow} E$
(b) $\sim C \mathbin{\Box\!\!\rightarrow} \sim E$

When such a pair of counterfactuals holds (in w) we can say that the presence of C *makes a difference* to the presence of E (in w).[1] It is important to stress how these two conditions play different roles: on the one hand (according to (a)) the presence of E is dependent on the presence of C, and on the other hand (according to (b)) the absence of E is dependent on the absence of C. Both of these conditions are necessary, and neither of them is trivial (for semantic considerations *cf.* Briggs 2012; List and Menzies 2009; Pernu 2016; Woodward 2003).

Now, armed with this notion of causation the causal exclusion problem can be tackled in the following way. Consider the following pair of counterfactuals:

(1a) $N \Box\!\!\rightarrow B$
(1b) $\sim N \Box\!\!\rightarrow \sim B$

And consider in comparison the following pair:

(2a) $M \Box\!\!\rightarrow B$
(2b) $\sim M \Box\!\!\rightarrow \sim B$

Which one of these pairs should we take to hold in the actual world? Supposing the basic assumption of nonreductive physicalism, namely the idea that mental states are not only realised, but multiple realised by underlying neural states, it becomes apparent in a fairly straight-forward way that it is actually the latter pair, rather than the former, that holds in the actual world. For suppose that M is present in the actual world, then even if M would happen to be realised by N, it could have equally well been realised by a different neural state, in which case (1b) becomes false while (2b) still holds: the absence of B is not dependent on the absence of N, for a different realiser of M could play the same exact role as N.

Although the difference-makers shun the use of the notion of causal sufficiency, their reasoning can actually be stated rather clearly by relying on the distinction between sufficiency and necessity. Call the above condition (a) the *sufficiency*, and the condition (b) the *necessity criterion* or *element* of causation. What we could now say is that although N is sufficient for B—that given that N is present, B will also be present—only M is also necessary—that given that M is absent, B will also be absent. This is the essence of the idea that it is mental states, rather than their contingent neural realisers, that make the difference to whether certain behaviour results or not. And this idea, it is claimed, can be empirically illustrated by relying on research on neuroprosthetics.

[1] That for a property C to cause the presence of another property E both (a) and (b) must hold entails that causation is proportional (in the vein of Yablo 1992). The proportionality constraint is not universally accepted, of course. In particular, interventionism (Woodward 2003) is not inherently committed to it. However, the proportionality constraint is in central role in the argumentation at hand, and accordingly the discussion in here is confined to the difference-making account of causation (although difference-making and interventionism are closely related doctrines—and they are conflated regrettably often—the former is a more stronger precisely in that it incorporates the proportionality constraint). To be clear, there are other counterarguments to the causal exclusion argument that are based on interventionism, but do not employ the proportionality constraint (e.g. Shapiro 2010; Shapiro and Sober 2007). They may be problematic in other ways (*cf.* Pernu 2013, 2014a, b), but they are not affected by the argumentation presented in here.

3 A crash course in neuroprosthetics

The roots of neuroprosthetics are in elementary neurophysiology. One of the major moments of scientific advancement occurred in 1849 when Hermann von Helmholtz carried out the first electrophysiological experiments and managed to measure, with a modified galvanometer, the speed at which the signal is propagated along a nerve fibre of a dissected frog (von Helmholtz 1850a, b). What's particularly important about this result, of course, is that it connected neural, and hence cognitive and motor processes to a recently born field of physics, namely electrodynamics. Helmholtz is arguably the most important single figure in the process of making us realise how mental phenomena can be accounted for in wholly physical terms (*cf*. Papineau 2001, 2002); the mind, as any physical entity, is spatiotemporally extended:

> Perhaps this idea is so familiar that it is difficult to get excited about it now. But it would be hard to exaggerate the impact of Helmholtz's findings when they were first announced. The nervous system is the 'organ of the mind', yes. But it is still a biological organ, and a physical one, subject to physical constraints. Its actions take time. (Mook 2004, p. 41.)

However, not only did Helmholtz's experiments give us solid evidence against Cartesian dualism. They also gave us very practical demonstration of the mechanism of motor control. In the experiments the nerve fibre was attached to the calf muscle (the whole nerve-muscle system had been dissected out from the frog and studied in isolation). When the nerve fibre was stimulated with an electric current, the muscle would contract (as Galvani's 1791, 1794 results had already indicated). This finding can be seen to contain the seeds of neuroprosthetics: the key to the generation and control of actions is in the electronic interface between the nervous system and muscles.

A lot has happened in neuroscience in the past 150 years, of course. Although techniques have naturally developed and changed, the basic idea of neuroprosthetics has remained the same: namely, harnessing the electrodynamical features of the nervous system, at different levels of biological organisation, and with better precision. Electroencephalography (EEG) was invented in the 1920s (Berger 1929), and is still tremendously useful in many areas of medicine and clinical work (e.g. determining seizure types, monitoring anaesthesia and the level of awarenesses, diagnosing brain death etc.). EEG-based (non-invasive/extra-cranial) neuroprosthetics have also been developed, and they have been used in helping paralysed or locked-in patients to communicate and operate devices (e.g. Birbaumer 2006; Birbaumer et al. 1999; Obermaier 2003; Sheikh et al. 2003; Wolpaw 2004). However, EEG is limited in spatiotemporal resolution, and thus generating and controlling more complicated and fine-grained action more sophisticated techniques must be used.

As technology advanced, it started to become possible to record the activity of single neurons and their ensembles, and it was demonstrated how monkeys could learn to control the firing rates of their cortical neurons (Fetz 1969; Fetz and Baker 1973; Fetz and Finocchio 1971, 1972, 1975). Soon it was suggested that extracting such data on neural activity from the motor cortex could be used to control the movements of external prosthetic devices (Schmidt et al. 1978; Schmidt 1980). However, a number of theoretical, methodological and technological obstacles prevented this from

becoming immediate reality. One important theoretical breakthrough, and one that is of a particular interest in the current context, was the discovery of a precise mathematical relationship between ensembles of cortical neurons and arm movements; more specifically, when individual neurons involved in a given motor task were represented as vectors, the direction of the resulting vector sum of all these individual vectors—*a neuronal population vector*—was found to be correlated with the direction of the arm movement (Georgopoulos et al. 1986, 1989). In other words, it was determined that there are specific neural correlates that are strikingly isomorphic to actual, external movements.

The basic neurophysiological elements of neuroprosthetics are now in place. A specific neural activation had been associated with motor tasks, and what remained was the engineering project of developing external devices that utilise these activations to generate functional behaviour. However, the neurophysiological research has also made notable advancements. One significant development has been the application of these general principles to cortical areas beyond the motor cortex. Research has been made on interfaces that connect to cognitive cortical signals, and on devices that respond to specific plans and intentions of motor tasks, rather than on the neural execution of those tasks (e.g. Hatsopoulos et al. 2004; Musallam et al. 2004; Shenoy et al. 2003). One could presume that devices functioning on the basis of conscious intentions would prove to be very useful indeed, for such intentions could be executed in a variety of ways, and they could thus allow a variety of different concrete implementations; indeed, the fruitfulness of this approach has recently been demonstrated on a human subject (Aflalo et al. 2015). However, it has also been suggested that an optimal system would rather be a hybrid one, consisting of interfaces connecting to both cognitive and motor cortical areas (Kim et al. 2006; Lebedev and Nicolelis 2009).

There are of course many empirical and technological details to attend to, and the field is developing rapidly. However, this concise rundown of the basic ideas and developments of neuroprosthetics gives us enough background to assess the philosophical arguments that are of interest in here.

4 Neuroprosthetics and mental causation

Although Helmholtz may have had some philosophical motives in carrying out his seminal research, the current work on neuroprosthetics is not trying to establish particular metaphysical theses; its goals are completely pragmatic. To assess the import of this research on the metaphysical debate concerning mental causation heavy-duty philosophical interpretation must thus be applied. There are two separate issues to address: first, what sort of relationship of the mental and the physical the research on neuroprosthetics imposes on us, and second, what sort of verdict on the debate over the causal efficacy of the mental can we expect to be delivered based on such a relationship?

To answer these question, let us start from the concrete arguments trying to ground mental causation on the research on neuroprosthetics (List and Menzies 2009; Menzies 2015; Woodward 2008a, b, 2017). These arguments run along the same lines, almost *in verbatim*. They are based on the most recent research that has focused on interfaces that

connect to cognitive cortical signals (Musallam et al. 2004 in particular), presumably because this work is explicitly engaged with intentional action—a notion that is of direct philosophical interest. The main point of these arguments is to draw attention to the fact that the results of this research are based on recordings of an ensemble of single neurons. There is thus some redundancy in neural activity at the level of single neurons; different patterns of neural activity at the level of single neurons could correspond to the same pattern of neural activity at the level of the ensemble. Now, although a particular pattern of neural activity at the level of single neurons is sufficient for generating a given action, there would always be different patterns that would be equally sufficient for the same exact action to be generated. But in that case it is not correct to say that had the given pattern of neural activity been absent, the action would also have been absent. So contrary what the causal exclusion argument seems to suggests, it is not so that the particular neural realiser of an intention should be deemed as the cause of the resulting action. Further, it now seems that it is actually the intention (rather than the neural realiser) that should be deemed the proper cause of the given action because couched in intentional terms the relevant counterfactual comes out true: had this particular intention been absent, the given action would also have been absent.

Here is a concrete example of this way of arguing:

> Suppose then that on some specific occasion t a monkey forms an intention I_1 to reach for a particular goal – call this action R_1. Suppose N_{11} is the particular (token) pattern of firing in the relevant set of neurons that realizes or encodes the intention I_1 on this particular occasion. Assume also that there are other token patterns of neural firing, N_{12}, N_{13} that realize the same intention I_1 on other occasions, so that I_1 is multiply realized by N_{11}, N_{12}, etc. The preference for micro or fine grained causation that we are considering recommends that we should regard N_{11} as the real cause of R_1 on occasion t. (Woodward 2008a, p. 239.)

> [T]he causal claim/causal explanation that appeals to N_{11} to explain R_1 seems overly specific. It fails to convey a relevant pattern of dependence: that there are some alternatives to N_{11} (namely, $N_{12}$and N_{13}) that would have led to the same reaching behavior R_1 and other alternatives (those that realize some different intention I_2, associated with reaching for a different goal) that would not have led to R_1. (Woodward 2008a, p. 239.)

The core of this argument rests on the idea that mental states (intentions) can be multiply realised. The multiple realisability thesis is widely shared, of course, and it forms the basis of nonreductive physicalism. Relying on the thesis is thus understandable, and many would think it is perfectly justifiable. However, that is not the case. The thesis is often used all too loosely in current philosophy, and this argument is a case in point. Intentions for generating particular actions are assumed to be multiply realised by different patterns of neural activity at the level of single neurons. Although this assumption is based on a perfectly correct understanding of one important aspect of neuroprosthetics, another important aspect has been neglected: the fact that these patterns of neural activity have something in common.

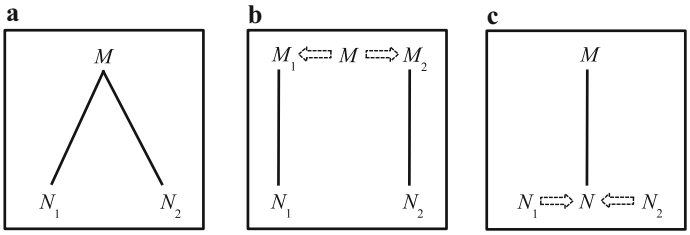

Fig. 1 **a** The multiple realisation hypothesis, **b** kind splitting, and **c** realiser unification/merging

Despite its popularity, the multiply realisability thesis has also been facing persistent criticism (e.g. Bechtel and Mundale 1999; Bickle 1998, 2003; Polger and Shapiro 2016; Shapiro 2000). One core aspect of the criticism can be stated as a dilemma: on the one hand, the realisers of the purportedly multiply realised entity must be identical, for they are all realising the same entity, in which case the realised entity is not differently realised after all; on the other hand, the realisers must differ from each other, for the realised entity should be differently realised, in which case there is no single entity to be multiply realised (*cf*. Couch 2004; Polger and Shapiro 2016; Shapiro 2000). Consequently, purported cases of multiple realisation typically dissolve in either of two ways: by realiser unification (merging) or kind splitting. In the former case the different realisers are recognised to share a feature with which the purportedly multiply realised entity is identified. In the latter case the different realisers are recognised to be fundamentally different, and the purportedly multiply realised entity is split into different entities (kinds). Figure 1 represents these different options.

Suppose now that M stands for the mental phenomenon we call "memory", and N_1 and N_2 stand for different neural realisations of this phenomenon. The claim is now, as depicted in Fig. 1a, that this phenomenon is multiple realised (*cf*. Craver 2004, 2007; Funkhouser 2014). This claim has intuitive credibility: we attribute this capacity to a large variety of biological organisms and artificial systems. However, when we look more closely into the neural basis of memory, the unity of the capacity starts to crumble: first, there is the fundamental distinction between short and long-term memory; second, the latter phenomenon is typically thought to split into two psychologically distinct phenomena, declarative and non-declarative (procedural) memory, and these in turn split further down into subcapacities, each of which are individuated by their neuroanatomical role (*cf*. e.g. Kandel and Pittenger 1999; Squire and Knowlton 1994; Schacter 1996; Thompson and Kim 1996). In consequence it is widely agreed in the neurosciences that there is no single, unitary phenomenon of memory; the notion of "memory" might be pragmatically useful to us, but in reality it encompasses a diversity of natural phenomena each with own distinct neural characteristics. So what we seem to be facing here is depicted in Fig. 1b: the higher-level mental phenomenon splitting into different kinds, each aligned with their neural realisers.

The case at hand illustrates the remaining option: the purported instance of multiple realisation dissolves into realiser unification or merging (Fig. 1c). To get a clearer picture on what is at issue here, consider the relationship between micro and macrophysics, between statistical mechanics and thermodynamics in particular. Now, one

could admit that there is some sort of multiple realisation between these two types of states since quite literally different microstates can realise the same macrostate. In fact, the very defining feature of the distinction is the fact that a large number of equiprobable microstates correspond to a particular macrostate, as witnessed by these passages from classic textbooks on statistical physics:

> [If] one deals with an isolated system, the *macro*state of the system might be specified by stating the values of the external parameters of the system (e.g., the value of the volume of the system) and the value of its constant total energy. The representative ensemble for the system is prepared in accordance with the specification of this macrostate; e.g., all systems in the ensemble are characterized by the given values of the external parameters and of the total energy. Of course, corresponding to this given *macro*state, the system can be in one of a very large number of possible *micro*states [...]. (Reif 1965, pp. 66–67.)

> The specification of the actual values of the parameters N [the number of identical particles], V [the volume] and E [the total energy] then defines the *macrostate* of the system. At the molecular level, however, a large number of possibilities still exist, because at that level there will *in general* be a large number of ways in which the macrostate (N, V, E) of the given system can be realized. (Pathria 1972, p. 10.)

Whether the relationship between macro and microstates in statistical mechanics is really of the sort that we should hold as "multiple realisation" is not of concern in here (no doubt Polger and Shapiro 2016, for example, would disagree with such a view on multiple realisation). What is essential is that an ensemble of values of higher-resolution variables can give rise to a single, specific value of a lower-resolution variable in a purely physical context; this sort of "multiple realisation" is consistent with reductive physicalism.

It now seems that the situation in neuroprosthetics is perfectly analogous to this. First, there is variety in neural activity at the level of single neurons; a particular intention or motor task is not associated with a single, specific distribution of neural activity at the level of single neurons. Second, however, each distribution of neural activity at the level of single neurons corresponds to single, specific neural activity at the level of the given ensemble of neurons. In other words, for each value of the mental variable of interest M, there exists a single, specific value of the neural variable N. What N is, however, is not a specific distribution neural activity at the level of single neurons—like the physical quantity that grounds a specific value of temperature is not identical to a specific distribution of molecular kinetic energies—but rather the average neural activity that the given neural ensemble gives rise to. And as N is determined by the average of the neural ensemble it is hardy surprising that different distributions of neural activity at the level of single neurons can give rise to the same value of N. What we could say, therefore, is that N is a "macroneural", rather than a "microneural" variable, but it is a neural variable nevertheless. And the mistake, it now becomes apparent, has been to assume that any given mental variable M—such as an intention to grasp an object—should be identified with a specific neural activity

at the level single neurons, such as N_1 or N_2, rather than with the neural feature that these specific states have in common, namely the average of their activity, N.

So, first, what sort of relationship of the mental and the physical does the research on neuroprosthetics impose on us? The results suggest that the mental and the physical are identical. Although there is variety in neural activity at the level single neurons, there are clearly identifiable physical features corresponding to different mental states. What are these features? They are the average firing rates of neurons in the relevant ensembles. In other words, there is a precise physical quantity corresponding to each intention. That is why those intentions are able to generate the given actions, and that is why neuroprosthetics enables us to connect devices on well-defined cortical areas that correspond to specific intentions. In fact, it is difficult to make sense of the whole research agenda unless one assumes that there are precise physical quantities to be defined and to connect to. A quote from Musallam et al. (2004) is telling: "if an individual has two potential reach goals, an apple and an orange, and the subject prefers apples over oranges, there are signals in his or her brain that indicate this preference and will influence the decision to reach for the apple instead of the orange" (p. 258). If there is a difference in intentions, there is a difference in the underlying neural activity, and this difference can be utilised in creating prosthetic devices.

Second, what sort of verdict on the debate over the causal efficacy of the mental can we now reach? It is true that we can now conclude that the mental is not an epiphenomenon: mental states (intentions) can be said to be causally efficacious. However, this is only so because the mental has now been identified with the physical. In other words, the causal efficacy of the mental has been saved at the price of the distinctness of the mental and the physical. So contrary to what the difference-making argumentation has been suggesting, the research on neuroprosthetics cannot be employed to save the idea of autonomous, distinct mental causation. Rather, neuroprosthetics is a particularly vivid demonstration of the fact how the mental is causally efficacious precisely because it is identical to the physical.

5 The distinctness of "the mental" and "the physical"

Reflecting on the research on neuroprosthetics, Woodward (2008a) points to the following conclusion: "From an interventionist perspective, this is about as clear a case of mental causation as one could imagine, since the subject uses the formation of one intention rather than another to manipulate the position of the limb" (p. 238; Woodward 2008b, pp. 163–164 makes exactly the same statement). It is worth examining in detail what is wrong—and what is right—about this claim.

What do we mean by "mental causation"? Or more precisely, since a difference-making account of causation has been assumed, what do we mean by "mental" (*cf.* Pernu 2017)? There is a distinct cortical—physical—variable corresponding to the formation of one intention rather than another; this is the basic idea of neuroprosthetics, and should not be at issue. Yes, these cortical-level variables (their specific values) correspond to various different neural activity at the level of single neurons, but it is this cortical (neural, physical) variable, and not the mental variable (intention), that is being thus "multiply realised". It is difficult to see how this should be "about as

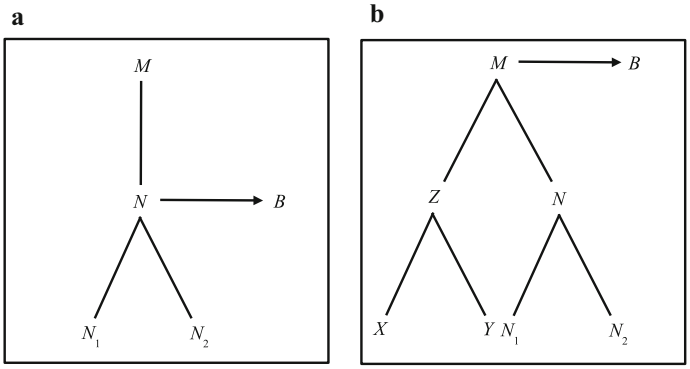

Fig. 2 **a** The case of neuroprosthetics where a mental state (intention) M is being realised by a cortical-level neural state N with two different, equally possible realisations (N_1 and N_2) at the level of single neurons; **b** the hypothetical case of the mental state M being multiply realised by two different cortical-level neural states, N and Z (and their lower-level realisers). B is the behavioural outcome that can be said to be dependent either on N or on M

clear a case of mental causation as one could imagine". On the contrary: this seems to be a rather clear case of physical (brain to bodily movement) causation. The problem is that the argument seems to locate the phenomenon of multiple realisation at the wrong place. The issue is phrased in terms of a mental state (intention) being multiply realised at the level of single neurons. But that is wrong, or thoroughly imprecise at least. What is being multiply realised (if this qualifies as multiple realisation in the first place) is the cortical-level neural states to which each of the intentions correspond. Each intention variable is identical with these cortical-level variables.

Consider Fig. 2a, b. Figure 2a represents the situation that we are actually faced with in neuroprosthetics: a mental state (intention) M is being realised by a cortical-level neural state N with different, equally possible distributions of neural activity at the level of single neurons, represented here by two possibilities, N_1 and N_2. Figure 2b represents a hypothetical scenario where the mental state M has two different potential cortical-level (or, more generally, coarse-grained) realisers (which are in turn are multiply realised at a lower-level). It is the situation depicted in Fig. 2b that those eager to offer "about as clear a case of mental causation as one could imagine" should be discussing. However, it is the situation depicted in Fig. 2a that is actually presented in the case of neuroprosthetics.

It is worth noting, however, that something closely akin to what is depicted in Fig. 2b has been in the focus of the traditional discussion. What Putnam (1967) and Fodor (1974), for example, originally stressed, when introducing the notion of multiple realisability to the modern debate in philosophy of mind, was that mental states could not be realised simply by a variety of different neural states, but that structurally very different creatures (in principle even extra-terrestrials and artificial intelligence) could legitimately be said to occupy the same mental states. According to this traditional view, the "Z" in Fig. 2b could therefore stand for widely different things. There are reasons to be sceptical of this view too, of course—the main doubt being that such purported cases of multiple realisation would be prone to be dissolved by kind splitting

(Fig. 1b)—but at least such a view seems closer to addressing the right issue. Note also that most of the research on neuroprosthetics has been done on monkeys. The research would not make sense, and the results could not have been transferred to practical therapeutic use, unless it would have been clear that the neural states corresponding to the different intentions would *not* be realised in different ways in different species. The success of neuroprosthetics could therefore be seen to speak against even the more traditional ideas of Putnam (1967) and Fodor (1974).

Maybe one could try to give a more charitable reading of the argumentation? Consider putting the issue in this way: is it right to place B (the behavioural outcome) "on the same level" with N (rather than with M, or somewhere in between)? This question is at the core of the issue, and it seems that it is not getting the attention it deserves. Of course one could now claim, as it has already been stressed, that M is simply identical with N, and that therefore neuroprosthetics provides us with "about as clear a case of mental causation as one could imagine" (and that B should be placed at the level of M, or somewhere in between the two). It should be rather evident, however, that that is not the intended reading. Identity is a symmetrical relation, and therefore that reading would commit one to also claiming that neuroprosthetics provides us with a clear case of physical causation. But of course the whole point of the debate is the contrast between the mental and the physical, and in granting mental states genuine causal power, one is typically tacitly saying that the physical (subvening) states lack that particular power. And even more to the point: the gist of the difference-making argumentation rests on the idea that it is the instantiation of mental properties *rather than* the instantiation of the subvening neural properties, that act as difference-makers and hence the proper causes of the behavioural outcomes. This contrastive result is achieved by assuming multiple realisability. If there is no such thing, the whole argumentation won't get off the ground (or you would end up with something that List and Menzies (2009) call the "compatibility result" – something that they don't see as providing a satisfactory solution to the problem of mental causation exactly because it does not grant the mental an autonomous role).

Here is another way of putting this point. One of the main premises of the causal exclusion argument is the distinctness of the mental and the physical. That is why the threat of epiphenomenalism is looming: if the mental and the physical are distinct, and the latter realm is causally complete, then the former seems thoroughly epiphenomenal (barring overdetermination). Of course you can get off this dire result by renouncing the distinctness of the mental and the physical, as Kim (1998, 2005) does. But if that is your solution, then it should not be advertised as breaking news. And, consequently, that is not what the difference-makers are after. What they want to show is that it is the mental states, and not the subvening neural states, that should be designated as the proper causes of the given behavioural outcomes—that it is the counterfactuals citing the instantiation of mental properties (counterfactuals of the type of 2b) rather than the counterfactuals citing the instantiation of neural properties (counterfactuals of the type of 1b) that come out true in these cases. And indeed, as List and Menzies (2009) show, this result follows neatly if you assume that causal relations involving mental properties are "realisation insensitive" (*vis-à-vis* their neural realisers): there are nearby possible worlds where the same mental properties are instantiated, and the same behavioural outcomes result, but where the actually instantiated neural properties

are absent. All this is fine as a formal result, of course. But what we need to ask—what we always need to ask when presented with formal models—is how, exactly, does any of this relate to the actual world? What and how, in this case, is "realisation insensitive"? Clearly the closest possible worlds are the ones where the cortical-level neural state N is being realised by different distributions of neural activity at the level of single neurons, e.g. N_1 and N_2. But all these worlds instantiate the same physical property, namely N, and in every relevant possible scenario where the intention is lacking, N is absent too, and hence the desired contrast between the two types of counterfactuals does not hold anymore. To get to the desired result you would need to move farther away, to scenarios where N would be absent but where M would still be instantiated (realised by Z rather than N say) and where the same behavioural outcome would still result. But nothing of that sort can be extracted from inspecting the research on neuroprosthetics. On the contrary: since the whole research paradigm is based on comparative neurophysiology, one can presume that one would need to travel quite far away indeed to encounter anything close to a scenario where M and B would hold but where N would be absent.

What this philosophical argumentation based on neuroprosthetics manages to make clear though, is that sometimes we need to point to coarse-grained variables to get to the right causal picture. And maybe it is a good thing to remind people of this. But surely this was never the main issue? What the difference-making argumentation can show is that sometimes macrophysical explanations can surpass microphysical explanations, but this is hardly surprising: we already knew—or at least had overwhelming reasons to suspect – that causation is a macrophysical phenomenon. Physicalism is not microphysicalism; no-one has thought (apart from Merricks 2001 perhaps) that physicalism leads us to eliminate the macrophysical world (*cf*. Hüttemann 2004; Hüttemann and Papineau 2005; Papineau 2013). The problems start to arise when the mental is assumed to be distinct from the physical, and when it is thought to be capable of injecting its own causal influence into the physical world. Neuroprosthetics is not suffering from such problems, however. The bodily, or robotic, movements—i.e. physical events—that we seek to control by implementing neuroprosthetic devises respond reliably and coherently to distinct macrophysical variables. Physical effects are fully accounted for by physical causes, and nothing in this research paradigm urges us to postulate any distinct mental variables.

6 Conclusion: Where do we go from here?

Relying on difference-making considerations have become an increasingly popular way of defending the causal efficacy of the mental. What has gone largely unnoticed, however, is that there are two different ways such considerations can be put into work for such a defence. On the one hand, one can hold that the mental and the physical are distinct, by relying on the multiple realisability thesis, and argue that since the counterfactuals concerning mental states, rather than counterfactuals concerning the underlying physical states, come out true, it is the mental, rather than its physical basis, that is genuinely causally efficacious. On the other hand, one can reject the multiple realisability thesis, and hold that the mental and the physical are identical,

but still argue that the mental is genuinely causally efficacious simply because mental causes *are* physical causes. It is this latter view, not the former, that the research on neuroprosthetics can be argued to give support to. It would therefore be immensely useful if the future discussion on mental causation would keep these two approaches carefully apart, and when the aim is to argue for the former, more substantial idea, great care should be exercised in formulating the multiple realisability thesis in a coherent and realistic manner.

Till now, however, the appeals to the research on neuroprosthetics have brought more confusion than clarity and precision to the discussion. This is very unfortunate. The neuroprosthetic framework provides a setting in which decision-making and action-control can be fruitfully and precisely studied. There is plenty of empirical data, produced by quantitative methods and phrased in terms of well-defined concepts. We should welcome this framework to philosophical discussion, for its more profound understanding could help us to articulate our philosophical theses in a more rigorous and concrete manner. In that way, perhaps, we could start taking steps towards understanding what mental causation really is.

Acknowledgements I want to thank Dr Nadine Elzein, Prof. Kristian Donner and the three anonymous referees of *Synthese* for helpful criticism and comments on various versions of this paper. This work has been financially supported by The Finnish Academy of Science and Letters and the Waldemar von Frenckell foundation.

References

Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., et al. (2015). Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science*, *348*, 906–910.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, *66*, 175–207.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*, 527–570.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht: Kluwer.

Birbaumer, N. (2006). Brain–computer-interface research: Coming of age. *Clinical Neurophysiology*, *117*, 479–483.

Birbaumer, N., et al. (1999). A spelling device for the paralysed. *Nature*, *398*, 297–298.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*, 139–166.

Clausen, J. (2008). Moving minds: Ethical aspects of neural motor prostheses. *Biotechnology Journal*, *3*, 1493–1501.

Clausen, J. (2011). Conceptual and ethical issues with brain–hardware interfaces. *Current Opinion in Psychiatry*, *24*, 495–501.

Couch, M. (2004). Discussion: A defense of bechtel and mundale. *Philosophy of Science*, *71*, 198–204.

Craver, C. F. (2004). Dissociable realization and kind splitting. *Philosophy of Science*, *71*, 960–971.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon Press.

Fetz, E. E. (1969). Operant conditioning of cortical unit activity. *Science*, *163*, 955–958.

Fetz, E. E., & Baker, M. A. (1973). Operantly conditioned patterns on precentral unit activity and correlated responses in adjacent cells and contralateral muscles. *Journal of Neurophysiology*, *36*, 179–204.

Fetz, E. E., & Finocchio, D. V. (1971). Operant conditioning of specific patterns of neural and muscular activity. *Science*, *174*, 431–435.
Fetz, E. E., & Finocchio, D. V. (1972). Operant conditioning of isolated activity in specific muscles and precentral cells. *Brain Research*, *40*, 19–23.
Fetz, E. E., & Finocchio, D. V. (1975). Correlations between activity of motor cortex cells and arm muscles during operantly conditioned response patterns. *Experimental Brain Research*, *23*, 217–240.
Fodor, J. A. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese*, *28*, 97–115.
Funkhouser, E. (2014). *The logical structure of kinds*. Oxford: Oxford University Press.
Galvani, L. (1791). De viribus electricitatis in motu musculari commentarius. *De Bononiensi Scientiarum et Artium Instituto atque Academia commentarii 7*, 363–418.
Galvani, L. (1794). *Dell'uso e dell'attività dell'arco conduttore nelle contrazioni dei muscoli*. Bologna: San Tommaso d'Aquino.
Georgopoulos, A. P., Lurito, J., Petrides, M., Schwartz, A., & Massey, J. (1989). Mental rotation of the neuronal population vector. *Science*, *243*, 234–236.
Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*, 1416–1419.
Glannon, W. (2016). Ethical issues in neuroprosthetics. *Journal of Neural Engineering*, *13*, 021002.
Hatsopoulos, N., et al. (2004). Decoding continuous and discrete motor behaviors using motor and premotor cortical ensembles. *Journal of Neurophysiology*, *92*, 1165–1174.
Hüttemann, A. (2004). *What's wrong with microphysicalism?*. London: Routledge.
Hüttemann, A., & Papineau, D. (2005). Physicalism decomposed. *Analysis*, *65*, 33–39.
Kandel, E. R., & Pittenger, C. (1999). The past, the future and the biology of memory storage. *Philosophical Transactions of the Royal Society of London. Series B Biological Sciences*, *354*, 2027–2052.
Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
Kim, J. (2005). *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.
Kim, H. K., Biggs, S., Schloerb, D., Carmena, J., Lebedev, M., Nicolelis, M., et al. (2006). Continuous shared control stabilizes reach and grasping with brain–machine interfaces. *IEEE Transactions on Biomedical Engineering*, *53*, 1164–1173.
Lebedev, M. A., & Nicolelis, M. A. L. (2009). Brain-machine interfaces: Past, present and future. *Trends in Neurosciences*, *29*, 536–546.
List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, *106*, 475–502.
Menzies, P. (2015). The causal closure argument is no threat to non-reductive physicalism. *Humana. Mente Journal of Philosophical Studies*, *29*, 21–46.
Merricks, T. (2001). *Objects and persons*. Oxford: Oxford University Press.
Mook, D. G. (2004). *Classic experiments in psychology*. Westport, CT: Greenwood Press.
Musallam, S., Corneil, B. D., Greger, B., Scherberger, H., & Andersen, R. A. (2004). Cognitive control signals for neural prosthetics. *Science*, *305*, 258–262.
Obermaier, B., et al. (2003). Virtual keyboard controlled by spontaneous EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*, 422–426.
Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents*. Cambridge: Cambridge University Press.
Papineau, D. (2002). *Thinking about consciousness*. Oxford: Oxford University Press.
Papineau, D. (2013). Causation is macroscopic but not irreducible. In S. C. Gibb & R. Ingthorsson (Eds.), *Mental causation and ontology*. Oxfrod: Oxford University Press.
Pathria, R. K. (1972). *Statistical mechanics*. Oxford: Butterworth-Heinemann Ltd.
Pernu, T. K. (2013). Does the interventionist notion of causation deliver us from the fear of epiphenomenalism? *International Studies in the Philosophy of Science*, *27*, 157–172.
Pernu, T. K. (2014a). Causal exclusion and multiple realizations. *Topoi*, *33*, 525–530.
Pernu, T. K. (2014b). Interventions on causal exclusion. *Philosophical Explorations*, *17*, 255–263.
Pernu, T. K. (2016). Causal exclusion and downward counterfactuals. *Erkenntnis*, *81*, 1031–1049.
Pernu, T. K. (2017). The five marks of the mental. *Frontiers in Psychology*, *8*, 1084.
Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford: Oxford University Press.
Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion*. Pittsburgh: University of Pittsburgh Press.

Reif, F. (1965). *Fundamentals of statistical and thermal physics*. New York, NY: McGraw-Hill.
Schacter, D. L. (1996). *Searching for memory: The brain, the mind, and the past*. New York, NY: Basic Books.
Schmidt, E. M. (1980). Single neuron recording from motor cortex as a possible source of signals for control of external devices. *Annals of Biomedical Engineering*, *8*, 339–349.
Schmidt, E. M., McIntosh, J. S., Durelli, L., & Bak, M. J. (1978). Fine control of operantly conditioned firing patterns of cortical neurons. *Experimental Neurology*, *61*, 349–369.
Shapiro, L. A. (2000). Multiple realizations. *Journal of Philosophy*, *97*, 635–654.
Shapiro, L. A. (2010). Lessons from causal exclusion. *Philosophy and Phenomenological Research*, *81*, 594–604.
Shapiro, L. A., & 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*. Pittsburgh: University of Pittsburgh Press.
Sheikh, H., et al. (2003). Electroencephalographic (EEG)-based communication: EEG control versus system performance in humans. *Neuroscience Letters*, *345*, 89–92.
Shenoy, K. V., et al. (2003). Neural prosthetic control signals from planactivity. *NeuroReport*, *14*, 591–596.
Squire, L. R., & Knowlton, B. J. (1994). Memory, hippocampus and brain systems. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 825–837). Cambridge, MA: The MIT Press.
Thompson, R. F., & Kim, J. J. (1996). Memory systems in the brain and localization of a memory. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 13438–13444.
von Helmholtz, H. (1850). Vorläufiger Bericht über die Fortpflanzungs-Geschwindigkeit der Nervenreizung. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, *1850*, 71–73.
von Helmholtz, H. (1850). Messungen über den zeitlichen Verlauf der Zuckung animalischer Muskeln und die Fortpflanzungsgeschwindigkeit der Reizung in den Nerven. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, *1850*, 276–364.
Wolpaw, J. R. (2004). Brain–computer interfaces (BCIs) for communication and control: A mini-review. *Supplements to Clinical Neurophysiology*, *57*, 607–613.
Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.
Woodward, J. F. (2008a). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reductive explanation and special science causation*. Oxford: Oxford University Press.
Woodward, J. F. (2008b). Cause and explanation in psychiatry: An interventionist perspective. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*. Baltimore: The Johns Hopkins University Press.
Woodward, J. F. (2017). Intervening in the exclusion argument. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference*. Oxford: Oxford University Press.
Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*, 245–280.